

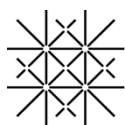
Intelligenzdiagnostik im Kindes- und Jugendalter:
Befunde zur Konstrukt-, differentiellen und Kriteriumsvalidität
der Intelligence and Development Scales

Inauguraldissertation
zur
Erlangung der Würde
einer Doktorin der Philosophie
vorgelegt der
Fakultät für Psychologie
der Universität Basel
von

Florine Schweizer

aus Basel, BS

Basel, 2018



**Universität
Basel**

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel edoc.unibas.ch

Genehmigt von der Fakultät für Psychologie

auf Antrag von

Prof. Dr. Alexander Grob

Prof. Dr. Sakari Lemola

Basel, den _____

Prof. Dr. Alexander Grob

DANKSAGUNG

Mein herzlichster Dank gilt folgenden Personen:

Prof. Dr. Alexander Grob—für die Möglichkeit, diese Dissertation zu schreiben, für die Unterstützung in den letzten vier Jahren und für das mir entgegengebrachte Vertrauen.

Prof. Dr. Sakari Lemola—für die Zweitbegutachtung meiner Dissertation.

Prof. Dr. Rui Mata—für den Vorsitz in meiner Promotionskommission.

Dr. habil. Priska Hagmann-von Arx—für die Inspiration, die unzähligen Inputs und die stets offene Tür für Fragen jeglicher Art.

Prof. Dr. Thomas Ledermann und Dr. Jasmin Gygi—für die vielen Inputs und die tolle Zusammenarbeit.

Dr. Rebekka Weidmann—für das Gegenlesen meiner Dissertation und die zahlreichen fachlichen und persönlichen Inputs.

Dr. Natalie Urfer, M.Sc. Anette Bünger und M.Sc. Jenna Wünsche sowie insbesondere M.Sc. Silvia Grieder und M.Sc. Salome Odermatt—für die diversen Inputs, die persönliche und motivationale Unterstützung und die stets offenen Ohren für Anliegen jeglicher Art.

Projektteam Testentwicklung—für die hingebungsvolle Unterstützung im Projekt und die unvergessliche Zeit.

Abteilungen PEP und EPP sowie SEED—für alle Inputs, kritischen Anmerkungen und Fragen in den Pausen, Coffee Breaks und D-Blocks.

Elisabeth Schweizer-Scherer und Leonhard Schweizer—für die bedingungslose Liebe und Unterstützung in allen Bereichen und den Glauben an mich.

Daniel Rubin—für alles, was man nicht in Worte fassen kann, und noch viel mehr.

INHALTSVERZEICHNIS

DANKSAGUNG	III
ABSTRACT	VI
1 Einleitung	1
2 Theoretischer Hintergrund	4
2.1 Testdiagnostik	4
2.2 Intelligenz	4
2.2.1 Intelligenztheorien	5
2.2.2 Intelligenzentwicklung	6
2.2.3 Intelligenzdiagnostik	6
2.2.4 Intelligence and Development Scales (IDS)	7
2.2.5 Weitere Testverfahren	10
2.3 Teststandards	10
2.3.1 Objektivität und Reliabilität	10
2.3.2 Validität	11
2.4 Validierungsbefunde der IDS	12
2.4.1 Konstruktvalidität	12
2.4.2 Differentielle Validität	14
2.4.3 Kriteriumsvalidität	16
3 Forschungsfragen	19
4 Methoden	21
4.1 Studien und Stichproben	21
4.2 Messinstrumente	22
4.3 Statistische Analysen	24
5 Resultate	26
5.1 Konstruktvalidität	26
5.2 Differentielle Validität	26
5.3 Kriteriumsvalidität: Prädiktive Validität	27
5.4 Kriteriumsvalidität: Konkurrente Validität	27
6 Diskussion	28
6.1 Konstruktvalidität	28
6.2 Differentielle Validität	29
6.3 Kriteriumsvalidität: Prädiktive Validität	31
6.4 Kriteriumsvalidität: Konkurrente Validität	32

6.5 Implikationen	33
6.6 Stärken und Limitationen	36
6.7 Ausblick	38
6.8 Konklusion	39
LITERATURVERZEICHNIS	40
APPENDIX A: Studie 1	54
APPENDIX B: Studie 2	99
APPENDIX C: Studie 3	116
APPENDIX D: Studie 4	126
APPENDIX E: Publikationsliste	148
APPENDIX F: Selbständigkeitserklärung	149

ABSTRACT

Die Intelligenz ist das am besten untersuchte und am häufigsten gemessene Merkmal der Psychologie und stellt einen wichtigen Prädiktor für zentrale Lebensbereiche dar. Unter anderem zeigen sich Zusammenhänge mit schulischen und beruflichen Leistungen, dem sozio-ökonomischen Status, Beziehungserfolg, Lebenszufriedenheit, Gesundheit sowie Langlebigkeit. Aus diesem Grund kommt Intelligenztests in der psychologischen Testdiagnostik eine große Bedeutung zu, beispielsweise bei Platzierungs- oder Fördermaßnahmenentscheidungen, zur Eignungsdiagnostik oder bei der Befunderhebung psychischer Störungen. Um mit Testverfahren gültige und generalisierbare Aussagen machen zu können, muss jedoch deren Validität gewährleistet sein. Diese kumulative Dissertation beschäftigt sich mit der Validierung der *Intelligence and Development Scales* (IDS). Die IDS liegen aktuell in drei deutschsprachigen Versionen vor: *Intelligence and Development Scales – Preschool* (IDS-P) für Kinder zwischen 3 und 5 Jahren, *Intelligence and Development Scales* (IDS-1) für Kinder zwischen 5 und 10 Jahren und *Intelligence and Development Scales – 2* (IDS-2) für Kinder und Jugendliche zwischen 5 und 20 Jahren. Anhand von vier Studien wird der Forschungsstand zur Konstrukt-, differentiellen und Kriteriumsvalidität der Intelligenz der IDS erweitert. Zum einen wurde die Konstruktvalidität der IDS-P und IDS-1 überprüft: Diese weisen faktorielle Validität im Sinne von Alters- und Geschlechtsinvarianz auf. Darüber hinaus bestehen nur wenige Geschlechtsdifferenzen in Untertests auf Mittelwertebene. Zum anderen wurde die differentielle Validität der IDS-2 hinsichtlich verschiedener Sprachgruppen untersucht: Die allgemeine Intelligenz, die Faktoren und die Untertests zeigen Mittelwertunterschiede zwischen Monolingualen, Bilingualen und Fremdsprachigen, in Abhängigkeit der Sprachgebundenheit der Skalen. Schließlich wurde die Kriteriumsvalidität der IDS-1 für Schulleistungen analysiert: Die allgemeine Intelligenz weist prädiktive Validität für Schulnoten über drei und über sieben Jahre auf. Des Weiteren ist die prädiktive Validität über drei Jahre mit derjenigen von anderen aktuellen deutschsprachigen Intelligenztests vergleichbar und Gewissenhaftigkeit sowie Leistungsmotivation zeigen zu Intelligenz inkrementelle konkurrente Validität. Zusammenfassend liefert die vorliegende Dissertation Evidenz, dass die IDS-P, IDS-1 und IDS-2 valide Testverfahren darstellen, die eine gültige Intelligenzdiagnostik im Kindes- und Jugendalter ermöglichen.

1 Einleitung

Die psychologische Testdiagnostik befasst sich mit der Einschätzung des Erlebens und Verhaltens von Personen anhand von Testverfahren (Beauducel & Leue, 2014; Ruch & Zimbardo, 2013). Intelligenztests erfassen dabei die Intelligenz—die Fähigkeit zum abstrakten und logischen Denken, zur Problemlösung, zum Planen und zum Lernen (Gottfredson, 1997). Die Intelligenz ist das am besten untersuchte Merkmal der Psychologie (Rost, 2009) und zeigt unter anderem positive Zusammenhänge mit Schulleistungen, Berufserfolg, dem sozioökonomischen Status, Beziehungserfolg, Lebenszufriedenheit, physischer und psychischer Gesundheit und Langlebigkeit sowie negative Zusammenhänge mit Substanzkonsum, Kriminalität und Unfällen (z.B. Aspara, Wittkowski & Luo, 2018; Calvin et al., 2011; Deary, 2009; Fergusson, Horwood & Ridder, 2005; Gonzalez-Mulé, Carter & Mount, 2017; Gottfredson & Deary, 2004; Roth et al., 2015; Schmidt & Hunter, 2004; Strenze, 2007; Wrulich et al., 2014). Intelligenztests sind folglich ein wichtiger Bestandteil der Diagnostik und gehören zu den am häufigsten eingesetzten Testverfahren (Goldstein, Princiotta & Naglieri, 2015). Die Intelligenzeinschätzung wird beispielsweise in der Schulpsychologie für Entscheidungen zur Beschulungsform oder zu Fördermaßnahmen eingesetzt, im beruflichen Bereich zur Eignungsdiagnostik oder zur Personalentwicklung herangezogen sowie in der klinischen Psychologie bei der Befunderhebung psychischer Störungen oder als Hinweis für weiterführende Abklärungen verwendet (Döpfner & Petermann, 2008; Holling, Preckel & Vock, 2004).

Um mit den verwendeten Testverfahren gültige Aussagen machen zu können, müssen diese verschiedenen Standards, sogenannten Testgütekriterien, entsprechen. Das wichtigste Kriterium stellt dabei die Validität dar, welche sich auf die Merkmalssättigung eines Tests bezieht und Aussagen bezüglich dessen Generalisierbarkeit erlaubt (Moosbrugger & Kelava, 2012). Dabei werden unter anderem drei Aspekte unterschieden: (1) Konstruktvalidität besteht, wenn vom Verhalten innerhalb der Testsituation auf ein zugrundeliegendes Konstrukt geschlossen werden kann. Ein Test sollte eine hohe Übereinstimmung mit Tests für gleiche oder ähnliche Merkmale (konvergente Validität) sowie eine vergleichsweise tiefe(re) Übereinstimmung mit Tests für andere Merkmale (divergente Validität) aufweisen; zudem sollte die Faktorenstruktur eines Tests (faktorielle Validität) geprüft werden. (2) Die differentielle Validität ermöglicht Aussagen darüber, inwieweit zwischen verschiedenen Subgruppen differenziert werden kann. (3) Kriteriumsvalidität liegt schließlich vor, wenn vom Verhalten innerhalb der Testsituation auf ein externes Kriterium geschlossen werden kann. Dabei kann ein Test ein zeitgleich erfasstes Kriterium querschnittlich (konkurrente Validität) oder längsschnittlich (prädiktive Validität) vorhersagen; zudem kann ein Zuwachs an Varianzaufklärung in einem Kriterium

resultieren (inkrementelle Validität), wenn ein Test zusätzlich zu einem anderen Test eingesetzt wird (Amelang & Schmidt-Atzert, 2006; Moosbrugger & Kelava, 2012; Schmidt-Atzert & Amelang, 2012).

Die vorliegende Dissertation untersucht diese drei genannten Validitätsaspekte der *Intelligence and Development Scales* (IDS) in Bezug auf die Intelligenz. Die IDS erfassen nebst der Intelligenz verschiedene Entwicklungsfunktionen und liegen in drei deutschsprachigen Versionen vor: *Intelligence and Development Scales – Preschool* (IDS-P; Grob, Reimann, Gut & Frischknecht, 2013) für drei- bis fünfjährige Kinder, *Intelligence and Development Scales* (IDS-1; Grob, Meyer & Hagmann-von Arx, 2009, 2013) für fünf- bis zehnjährige Kinder und *Intelligence and Development Scales – 2* (IDS-2; Grob & Hagmann-von Arx, 2018) für fünf- bis zwanzigjährige Kinder und Jugendliche¹. Während für die IDS-P, IDS-1 und IDS-2 verschiedene Evidenz zur Konstrukt-, differentiellen und Kriteriumsvalidität vorliegt (z.B. Grob & Hagmann-von Arx, 2018; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013), sind beispielsweise Befunde zur Messinvarianz und zu Geschlechtsdifferenzen für die IDS-P und IDS-1, Befunde zur Differenzierungsfähigkeit zwischen verschiedenen Sprachgruppen für die IDS-2 sowie Befunde zur Vorhersage von Schulleistungen durch die IDS-1 im Vergleich zu anderen aktuellen Testverfahren und im Zusammenhang mit weiteren relevanten Prädiktoren ausstehend.

Das Ziel dieser Dissertation ist es, diese Forschungslücken im Bereich der Intelligenz zu schließen und damit einen Beitrag zur Erweiterung der Validität und zur Generalisierbarkeit der IDS zu leisten. Im Rahmen der Konstruktvalidität werden die faktorielle Validität im Sinne der Messinvarianz über das Alter und Geschlecht sowie Geschlechtsdifferenzen auf Mittelwert- und Varianzebene in den IDS-P und IDS-1 untersucht (Studie 1). Im Rahmen der differentiellen Validität werden Mittelwertunterschiede zwischen Monolingualen, Bilingualen und Fremdsprachigen in den IDS-2 überprüft (Studie 2). Schließlich wird im Rahmen der Kriteriumsvalidität die prädiktive Validität der IDS-1 für Schulleistungen über drei (Studie 3) und über sieben Jahre (Studie 4) analysiert. Dabei wird die prädiktive Validität über drei Jahre derjenigen von anderen deutschsprachigen Intelligenztests—*Wechsler Intelligence Scale for Children – Vierte Version* (WISC-IV; Petermann & Petermann, 2011), *Reynolds Intellectual Assessment Scales* (RIAS; Hagmann-von Arx & Grob, 2014) und *Snijders-Oomen nonverbaler Intelligenztest 6-40* (SON-R 6-40; Tellegen, Laros & Petermann, 2012)—gegenübergestellt (Studie 3). Darüber

¹ Um in der vorliegenden Dissertation die verschiedenen IDS-Versionen voneinander abgrenzen zu können, wird im weiteren Verlauf der Arbeit der Ausdruck *IDS* für alle IDS-Versionen verwendet. *IDS-P* bezieht sich auf die Version für Drei- bis Fünfjährige, *IDS-1* auf die Version für Fünf- bis Zehnjährige und *IDS-2* auf die Version für Fünf- bis Zwanzigjährige. Zu den *IDS-1* ist anzumerken, dass diese in der Originalreferenz als *IDS* (und nicht *IDS-1*) bezeichnet werden.

hinaus wird der Frage nachgegangen, ob Gewissenhaftigkeit und Leistungsmotivation zusätzlich zur Intelligenz inkrementelle konkurrente Validität für Schulleistungen aufweisen (Studie 4).

Die vorliegende Dissertation ist wie folgt aufgebaut: Kapitel 2 fasst den theoretischen Hintergrund zusammen. Zuerst wird der Begriff Testdiagnostik erörtert, danach werden die Intelligenz, aktuelle Intelligenztheorien, die Entwicklung und Diagnostik der Intelligenz sowie die IDS und weitere für diese Dissertation relevante Intelligenztestverfahren näher beleuchtet. Im Anschluss werden Teststandards im Hinblick auf die Gütekriterien Objektivität und Reliabilität sowie insbesondere Validität beschrieben, bevor als Letztes Validierungsbefunde und Forschungslücken der IDS aufgezeigt werden. In Kapitel 3 werden die Forschungsfragen dieser Dissertation aufgeführt. Kapitel 4 beschreibt die Methoden und Kapitel 5 fasst die Resultate zusammen. Kapitel 6 schließt mit einer Diskussion hinsichtlich der zentralen Befunde, Implikationen für die Praxis, Stärken und Limitationen, einem Ausblick für zukünftige Forschung sowie einer zusammenfassenden Konklusion.

2 Theoretischer Hintergrund

2.1 Testdiagnostik

Die Psychologie ist die Wissenschaft des Erlebens und Verhaltens des Menschen über die Lebensspanne (z.B. Ruch & Zimbardo, 2013). Um das Erleben und Verhalten fassbar zu machen, bedient sich die psychologische Testdiagnostik verschiedener Testverfahren (Testkuratorium, 2010). Das Ziel besteht darin, Defizite und Ressourcen zu erkennen. Bei Defiziten sollte eine spezifische Intervention stattfinden, wobei Ressourcen gezielt genutzt und gefördert werden sollten (Grob, Hagmann-von Arx & Bodmer, 2009; Kaufman, Raiford & Coalson, 2016):

(. . .) Diagnostik beinhaltet die empirisch basierte, möglichst genaue Schätzung der Ausprägung und Veränderung psychologischer Konstrukte bei Merkmalsträgern sowie die möglichst genaue Klassifikation der Merkmalsträger in Gruppen mit ähnlichen psychischen Merkmalen, unter Beachtung transparenter, wissenschaftlicher und ethischer Standards sowie einer kompetenzbasierten, theorie- bzw. regelgeleiteten Integration und Interpretation der Informationen (Gutachten, Diagnose, Prognose) mit dem Ziel der Beantwortung diagnostischer Fragestellungen (z.B. Vorbereitung von Interventionsmaßnahmen und Entscheidungen). (Beauducel & Leue, 2014, S. 21)

Die Diagnostik unterscheidet verschiedene Arten von Testverfahren wie beispielsweise projektive Tests, Persönlichkeitstests und Leistungstests (Brähler, Holling, Leutner & Petermann, 2002). Letztere erfassen die kognitive Leistungsfähigkeit, entweder anhand von Geschwindigkeitstests (einfache Aufgaben mit begrenzter Bearbeitungszeit) oder Niveautests (schwierige Aufgaben mit unbegrenzter Bearbeitungszeit) oder als Mischform (Jonkisz, Moosbrugger & Brandt, 2012). Nebst Aufmerksamkeits-, Entwicklungs- und Schultests gehören hierzu Intelligenztests (Amelang & Schmidt-Atzert, 2006), die im Fokus der vorliegenden Dissertation stehen.

2.2 Intelligenz

Intelligenz ist ein Subaspekt der Persönlichkeit (Roberts & Wood, 2006), wird unter kognitive Funktionen subsumiert (Diamond, 2013) und beschreibt das „can do“ (Gottfredson, 2002, S. 37) sowie die „Fähigkeit, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156). Während Definitionen lange Zeit durch Uneinheitlichkeit geprägt waren (Pawlik, 1968), besteht heute ein Konsens darüber, was Intelligenz umfasst (Stern & Neubauer, 2016):

Intelligenz ist eine sehr allgemeine geistige Kapazität, die—unter anderem—die Fähigkeit zum schlussfolgernden Denken, zum Planen, zur Problemlösung, zum abstrakten

Denken, zum Verständnis komplexer Ideen, zum schnellen Lernen und zum Lernen aus Erfahrung umfasst. Es ist nicht reines Bücherwissen, keine enge akademische Spezialbegabung, keine Testerfahrung. Vielmehr reflektiert Intelligenz ein breiteres und tieferes Vermögen, unsere Umwelt zu verstehen, ‚zu kapieren‘, ‚Sinn in Dingen zu erkennen‘ oder ‚herauszubekommen‘, was zu tun ist. (Gottfredson, 1997, S. 13; deutsche Übersetzung in Rost, 2009, S. 2)

2.2.1 Intelligenztheorien. Um die Intelligenz als Konstrukt zu verstehen, wurden in der Vergangenheit verschiedene Theorien postuliert. Die aktuellste und umfassendste Theorie ist die Cattell-Horn-Carroll-Theorie von McGrew (CHC-Theorie; McGrew, 1997), welche eine Integration der Theorien von Spearman (1904), Cattell (1941), Horn (1965), Horn und Cattell (1966) sowie Carroll (1993) darstellt. Die CHC-Theorie besagt, dass die Intelligenz hierarchisch aufgebaut ist (Spearman, 1904) und aus drei Schichten (Strata) besteht (Carroll, 1993). Stratum III an der Spitze umfasst die allgemeine Intelligenz *g* (Spearman, 1904), darunter auf Stratum II befinden sich zehn Sekundärfaktoren (breite Fähigkeiten; Cattell, 1941; Horn, 1965; Horn & Cattell, 1966) und zuunterst auf Stratum I liegen über 70 Primärfaktoren (enge Fähigkeiten; McGrew, 1997). Folgende sind die Sekundärfaktoren auf Stratum II, geordnet nach ihrer Relevanz für *g*: *Fluide Intelligenz*² (*Gf*; genetisch determinierte Grundintelligenz und Fähigkeit für logisches und schlussfolgerndes Denken), *Quantitatives Wissen* (*Gq*; Fähigkeit zur Manipulation von quantitativen Informationen und numerischen Symbolen), *Kristalline Intelligenz*³ (*Gc*; sprach- und kulturgebundene Wissensaneignung und -anwendung), *Lesen und Schreiben* (*Grw*; Fähigkeiten des Lesens und Schreibens), *Kurzzeitgedächtnis* (*Gsm*; Verarbeitung, Speicherung und Abruf von Informationen im und aus dem Kurzzeitgedächtnis), *Visuelle Wahrnehmung* (*Gv*; Aufnahme, Speicherung und Abruf von visuellen Stimuli), *Auditive Wahrnehmung* (*Ga*; Aufnahme, Speicherung und Abruf von auditiven Stimuli), *Langzeitgedächtnis* (*Glr*; Verarbeitung, Speicherung und Abruf von Informationen im und aus dem Langzeitgedächtnis), *Verarbeitungsgeschwindigkeit* (*Gs*; automatische und flüssige Ausführung von kognitiven Aufgaben) und *Reaktions-/Entscheidungszeit* (*Gt*; Geschwindigkeit, mit der jemand auf Reize reagiert und Entscheidungen trifft; vgl. McGrew, 1997). Auf Stratum I stellen beispielsweise die Primärfaktoren *Induktion* und *Quantitatives Schlussfolgern* Teilleistungen von *Gf* dar, welche in der Regel direkt anhand von Untertests erhoben werden (McGrew, 2009). Um eine Einschätzung eines Faktors vornehmen zu können, benötigt es jeweils mindestens zwei Untertests (McGrew, 1997).

² Gemäß Baltes (1990) *Fluide Mechanik*

³ Gemäß Baltes (1990) *Kristalline Pragmatik*

2.2.2 Intelligenzentwicklung. Die Intelligenz entwickelt sich in Abhängigkeit von genetischen und Umweltfaktoren über die Lebensspanne. Insgesamt weist die Intelligenz eine relativ hohe Stabilität auf, sodass ab Schuleintritt Leistungsvorhersagen für die Grundschulzeit und ab etwa zehn Jahren längerfristige Prognosen gestellt werden können (Rost, 2009). Dabei werden drei Arten von Stabilität unterschieden (Rost, 2009): Die Strukturstabilität (auch Konstruktstabilität) bezeichnet den Aufbau eines Merkmals. Die Intelligenzentwicklung folgt in den frühen Lebensjahren einer Differenzierung (Ausbildung von Faktoren) und im höheren Alter einer Dedifferenzierung (Zunahme der Wichtigkeit von g ; Rost, 2009). Die Mittelwertstabilität (auch Niveau-, absolute oder normative Stabilität) bezeichnet die intraindividuelle Ausprägung eines Merkmals. Die Intelligenz zeigt eine kontinuierliche Zunahme ab Geburt mit einer Anstiegsabflachung mit zunehmendem Alter, eine Plateaubildung im Alter von ungefähr 25 Jahren und eine physiologisch bedingte Abnahme im höheren Alter (Schaie, 1994). Die Positionsstabilität (auch Rangreihen- oder differentielle Stabilität) bezeichnet die interindividuelle Rangordnung eines Merkmals. Die Intelligenz weist ab ungefähr zehn Jahren eine hohe Positionsstabilität auf (Deary, Whalley, Lemmon, Crawford & Starr, 2000).

2.2.3 Intelligenzdiagnostik. Intelligenz ist ein latentes, nicht direkt beobachtbares Konstrukt und wird mittels sogenanntem intelligenten Verhalten operationalisiert. Dieses wird in verschiedenen Situationen mit Intelligenztests erhoben und zusammenfassend als Intelligenzeinschätzung anhand eines Intelligenzquotienten (IQ) angegeben (Schneider & Flanagan, 2015). Der IQ einer Person zeigt auf, ob ihre Intelligenzausprägung im Normbereich liegt. Anhand einzelner Subbereiche können zusätzlich inter- und/oder intraindividuelle Stärken und/oder Schwächen abgelesen werden (Grob & Hagmann-von Arx, 2018). Der IQ der Population ist normalverteilt und besitzt in der Regel eine Skala mit einem Mittelwert von 100 und einer Standardabweichung von 15. Folglich liegt der IQ bei 68% der Personen zwischen 85 und 115, bei 95% zwischen 70 und 130 und bei über 99% zwischen 55 und 145 (Rost, 2009). Ein IQ unter 70 wird als Intelligenzminderung oder geistige Behinderung und ein IQ zwischen 70 und 84 als Lernbehinderung bezeichnet. Eine überdurchschnittliche Intelligenz liegt ab einem IQ von 116 und eine Hochbegabung ab einem IQ von 130 vor (Grob & Hagmann-von Arx, 2012).

Intelligenztests gehören zu den am häufigsten eingesetzten Testverfahren in der psychologischen Forschung und Praxis (Goldstein et al., 2015). Der erste Test wurde bereits vor über 100 Jahren entwickelt, um individuelle Unterschiede im Schulerfolg vorhersagen zu können (Binet & Simon, 1904). Mittlerweile sind eine Vielzahl an Tests auf dem Markt verfügbar, welche sich hinsichtlich Altersspanne, Umfang, Inhalt und theoretischem Hintergrund unterscheiden (z.B. Hagmann-von Arx, Meyer & Grob, 2008b). Die vorliegende Dissertation fokus-

siert auf die IDS (IDS-P: Grob, Reimann et al., 2013; IDS-1: Grob, Meyer et al., 2009, 2013; IDS-2: Grob & Hagmann-von Arx, 2018). Zusätzlich sind für die Dissertation die deutschen Versionen folgender drei Intelligenztests relevant: WISC-IV (Petermann & Petermann, 2011), RIAS (Hagmann-von Arx & Grob, 2014) und SON-R 6-40 (Tellegen et al., 2012).

2.2.4 Intelligence and Development Scales (IDS). Die IDS wurden in Anlehnung an den *Kramer-Intelligenztest* (Kramer, 1972) mit dem Ziel entwickelt, „das Kind in seiner Ganzheit [zu] verstehen und [zu] erfassen“ (Grob, 2018, S. 2). Aus diesem Grund ermöglichen sie als eines der ersten Testverfahren eine umfassende Einschätzung sowohl der Intelligenz als auch weiterer entwicklungsrelevanter Funktionsbereiche. Die IDS liegen in drei deutschsprachigen Versionen vor: IDS-P (Grob, Reimann et al., 2013), IDS-1 (Grob, Meyer et al., 2009, 2013) und IDS-2 (Grob & Hagmann-von Arx, 2018). Die IDS-1 für Kinder zwischen 5;0 und 10;11 Jahren stellen die erste Fassung dar und wurden als IDS-P und IDS-2 weiterentwickelt. Die IDS-P für das Vorschulalter (3;0 – 5;11 Jahre) erweitern den Altersbereich der IDS-1 auf jüngere Kinder. Die IDS-2 für das Kindes- und Jugendalter (5;0 – 20;11 Jahre) erweitern den Altersbereich der IDS-1 auf Jugendliche; zudem enthalten sie einige theoretische und inhaltliche Neuerungen gegenüber den IDS-1 und werden diese in naher Zukunft ablösen.

Die IDS finden unter anderem Anwendung in der Entwicklungs- und Schulpsychologie, in der Berufs- und Laufbahnberatung, in der Erziehungs- und Familienberatung, in der Sonder- und Heilpädagogik, in der klinischen Kinder- und Jugendpsychologie respektive -psychiatrie sowie in der Pädiatrie. Sie können zur Bestimmung von Begabungen, Schulart und Schulniveau sowie für eine Standortbestimmung im Kontext von Entwicklungs-, Verhaltens- oder emotionalen Störungen eingesetzt werden (Grob & Hagmann-von Arx, 2018; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). Aufgrund der positiven Resonanz der IDS wurden die IDS-P zudem in Polen (Fecenec, Jaworowska & Matczak, 2015) und die IDS-1 in Polen (Jaworowska, Matczak & Fecenec, 2011), Tschechien (Krejčířová, Urbánek, Širůček & Jabůrek, 2013) und Italien (Ferri, Rea & Casagrande, 2015) adaptiert. Für die IDS-2 sind Adaptationen in Polen (Fecenec, Jaworowska & Matczak, in prep.), Italien (Ferri, Casagrande & Rea, in prep.), England (Barnett & Stuart, in prep.) und den Niederlanden (Ruiter, Geerds, Timmerman & Visser, in prep.) in Arbeit sowie in Brasilien, Dänemark, Finnland, Frankreich, Norwegen, Schweden, Spanien und Tschechien in Planung.

Die IDS-P und IDS-1 erfassen die Intelligenz als Funktionsbereich *Kognition* im Bereich *Kognitive Entwicklung*. Die allgemeine Intelligenz (*IQ*) wird in Anlehnung an Spearman (1904) sowie Baltes (1990) anhand von sieben Untertests eingeschätzt (siehe Tabelle 1; vgl. Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). Die IDS-2 erfassen die Intelligenz als

Funktionsbereich *Intelligenz* im Bereich *Kognitive Funktionen*. Die allgemeine Intelligenz kann in Anlehnung an Spearman (1904) und die CHC-Theorie (McGrew, 1997) als *IQ-Profil* (7 Faktoren und 14 Untertests), *IQ* (7 Untertests) und/oder *IQ-Screening* (2 Untertests) anhand von sieben Faktoren mit jeweils zwei aufgabentypähnlichen Untertests eingeschätzt werden (siehe Tabelle 2; vgl. Grob & Hagmann-von Arx, 2018).

Tabelle 1

Intelligenzuntertests der IDS-P und IDS-1 (Funktionsbereich Kognition)

Untertest	Aufgabe	CHC-Faktor ^a
Wahrnehmung Visuell	Serien mit Abbildungen unterschiedlich langer Buntstifte (IDS-P) respektive Serien unterschiedlich langer Striche (IDS-1) bilden	Gv
Aufmerksamkeit Selektiv	Enten mit bestimmten Merkmalen aus einer Menge verschiedener Enten möglichst schnell und richtig aussortieren (IDS-P) respektive durchstreichen (IDS-1)	Gs
Gedächtnis Phonologisch	Kunstwörter ansteigender Silbenlänge (IDS-P) respektive alternierend Zahlen- und Buchstabenfolgen (IDS-1) unmittelbar und in derselben Reihenfolge wiedergeben	Gsm ^b
Gedächtnis Räumlich-Visuell	Figuren merken und aus einer Auswahl die Figuren wiedererkennen	Gsm ^c
Gedächtnis Auditiv	Eine semantisch sinnvolle Geschichte hören und nach mindestens 20 Minuten frei und gestützt wiedergeben	Glr
Denken Bildlich	Geometrische Figuren mit Hilfe von Rechtecken und Dreiecken nachlegen	Gf und Gv
Denken Konzeptuell	Bei drei Bildern das Gemeinsame (das Konzept) erkennen und dazu aus fünf weiteren Bildern zwei passende auswählen	Gf und Gc

Anmerkung. In Anlehnung an Grob, Meyer et al. (2013, S. 18; Testmanual) und Grob, Reimann et al. (2013, S. 20; Testmanual). *IQ*: alle 7 Untertests.

^a Die Untertests der IDS-P und IDS-1 erfassen nicht explizit CHC-Faktoren, lassen sich diesen respektive Teilaspekten davon jedoch zuordnen.

^b Phonologischer Arbeitsspeicher (nach Baddeley, 2003).

^c Räumlich-visueller Arbeitsspeicher (nach Baddeley, 2003).

Zusätzlich zur Intelligenz erfassen die IDS verschiedene Funktionsbereiche im Bereich *Allgemeine Entwicklung*: Alle IDS erfassen *Psychomotorik* und *Sozial-Emotionale Kompetenz*. Die IDS-P und IDS-1 erfassen zudem *Mathematik* und *Sprache*. Die IDS-2 erfassen *Schulische Kompetenzen* (Untertests *Logisch-Mathematisches Denken*, *Sprachliche Fähigkeiten*, *Lesen* und *Rechtschreiben*) und *Arbeitshaltung*; weiter wird im Bereich *Kognitive Funktionen* der Funktionsbereich *Exekutive Funktionen* eingeschätzt. Schließlich erfassen die IDS-P den *Umgang mit der Testsituation*, die IDS-1 die *Leistungsmotivation* und die IDS-2 die *Mitarbeit wäh-*

Tabelle 2

Intelligenzfaktoren und -untertests der IDS-2 (Funktionsbereich Intelligenz)

Faktor	Untertest	Aufgabe	CHC-Faktor
Verarbeitung Visuell	Figuren nachlegen	Geometrische Figuren mit Hilfe von Rechtecken und Dreiecken nachlegen	Gv
	Plättchen legen	Plättchen nach Vorlage nachlegen	
Verarbeitungs- geschwindigkeit	Zwei Merkmale durchstreichen	Papageien mit bestimmten Merkmalen aus Reihen verschiedener Papageien möglichst schnell und richtig durchstreichen	Gs
	Kästchen durchstreichen	Gruppen mit drei oder vier Kästchen aus Reihen verschieden großer Gruppen möglichst schnell und richtig durchstreichen	
Kurzzeit- gedächtnis Auditiv	Zahlen- und Buch- stabenreihen nach- sprechen	Zahlen- und Buchstabenreihen unmittelbar vorwärts und rückwärts nachsprechen	Gsm ^a
	Gemischte Zahlen- und Buchstabenrei- hen nachsprechen	Gemischte Zahlen- und Buchstabenreihen unmittel- bar vorwärts und rückwärts nachsprechen	
Kurzzeit- gedächtnis Räumlich- Visuell	Figuren wiedererkennen	Figuren merken und aus einer Auswahl die Formen und Positionen wiedererkennen	Gsm ^b
	Rotierte Figuren wiedererkennen	Figuren merken und aus einer Auswahl rotierter Fi- guren die Formen und Positionen wiedererkennen	
Langzeit- gedächtnis	Geschichte nacherzählen	Eine semantisch sinnvolle Geschichte hören und nach mindestens 20 Minuten frei und gestützt wie- dergeben	Glr
	Bild beschreiben	Ein Bild betrachten und nach mindestens 20 Minu- ten gestützt beschreiben	
Denken Abstrakt	Matrizen ergänzen	Erkennen, wie sich eine Figur verändert, und diese Veränderung auf eine weitere Figur anwenden	Gf
	Unpassende Bilder ausschließen	Aus einer Anzahl von Bildern entscheiden, welches Bild nicht dazu passt	
Denken Verbal	Kategorien nennen	Für eine Gruppe von Bildern oder Begriffen die korrekte Kategorie nennen	Gc
	Gegenteile nennen	Gegenteile zu vorgegebenen Wörtern nennen	

Anmerkung. In Anlehnung an Grob und Hagmann-von Arx (2018, S. 15; Testmanual zu Theorie, Interpretation und Gütekriterien). *IQ-Profil:* alle 14 Untertests; *IQ:* 7 Untertests (*Figuren nachlegen, Zwei Merkmale durchstreichen, Zahlen- und Buchstabenreihen nachsprechen, Figuren wiedererkennen, Geschichte nacherzählen, Matrizen ergänzen, Kategorien nennen*); *IQ-Screening:* 2 Untertests (*Matrizen ergänzen, Kategorien nennen*).

^a Phonologischer Arbeitsspeicher (nach Baddeley, 2003).

^b Räumlich-visueller Arbeitsspeicher (nach Baddeley, 2003).

rend der Testsituation anhand eines Testleiterfragebogens. Die Funktionsbereiche *Mathematik, Sprache, Schulische Kompetenzen* sowie *Exekutive Funktionen* beinhalten dabei kognitive Anteile (vgl. Grob & Hagmann-von Arx, 2018; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013).

2.2.5 Weitere Testverfahren. Die WISC-IV (Petermann & Petermann, 2011) entspringt der langjährigen Tradition der *Wechsler-Skalen* und ist für Kinder und Jugendliche zwischen 6;0 und 16;11 Jahren konzipiert. Sie erfasst in Anlehnung an die CHC-Theorie (McGrew, 1997) die allgemeine Intelligenz (*Gesamt-IQ*) auf Stratum III, die vier Indizes *Sprachverständnis, Verarbeitungsgeschwindigkeit, Wahrnehmungsgebundenes Logisches Denken* und *Arbeitsgedächtnis* auf Stratum II sowie 10 bis 15 Untertests auf Stratum I.⁴

Die RIAS (Hagmann-von Arx & Grob, 2014) stellen eine Adaptation der amerikanischen RIAS (Reynolds & Kamphaus, 2003) dar und sind für Personen zwischen 3;0 und 99;11 Jahren konzipiert. Sie erfassen in Anlehnung an Horn und Cattell (1966) die allgemeine Intelligenz (*Gesamtintelligenz Index*) auf Stratum III, die beiden Indizes *Verbaler Intelligenz Index* und *Nonverbaler Intelligenz Index* auf Stratum II sowie vier Untertests auf Stratum I; zusätzlich kann ein *Gesamtgedächtnis Index* anhand von zwei Untertests eingeschätzt werden.

Der SON-R 6-40 (Tellegen et al., 2012) gehört zu den *Snijders-Oomen nonverbalen Intelligenztests* und ist für Personen zwischen 6;0 und 40;11 Jahren konzipiert. Er erfasst die allgemeine Intelligenz (*SON-IQ*) im Sinne der fluiden Intelligenz (Cattell, 1941) anhand von vier Untertests.

2.3 Teststandards

Um mit Testverfahren gültige Schlüsse ziehen zu können, müssen diese verschiedenen testtheoretischen Standards genügen. Im deutschsprachigen Raum überprüft das Diagnostik- und Testkuratorium anhand eines Testbeurteilungssystems, ob die *Standards für pädagogisches und psychologisches Testen* (Häcker, Leutner & Amelang, 1998) und die *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen* (DIN 33430; Deutsches Institut für Normierung, 2002) eingehalten werden. Diese Standards und Anforderungen enthalten unter anderem die Haupttestgütekriterien Objektivität, Reliabilität und Validität (Testkuratorium, 2010; vgl. auch Kersting, 2006; Moosbrugger & Höfling, 2006, 2012).

2.3.1 Objektivität und Reliabilität. Objektivität ist die Standardisierung und Unabhängigkeit eines Tests von Testbedingungen hinsichtlich Durchführung, Auswertung und Interpretation (Moosbrugger & Kelava, 2012). Reliabilität bezeichnet die Zuverlässigkeit und Messge-

⁴ Mittlerweile ist die *Wechsler Intelligence Scale for Children – Fünfte Version* (WISC-V; Petermann, 2017) erschienen, die vorliegende Dissertation fokussiert jedoch auf die WISC-IV.

naugigkeit eines Tests und gibt den Anteil der wahren Varianz an der Gesamtvarianz der Testwerte an (Moosbrugger & Kelava, 2012). Für bedeutsame Entscheidungen auf Individualebene werden Koeffizienten $> .90$ als gut, $> .80$ als ausreichend und $< .80$ als ungenügend klassifiziert, während für Forschungszwecke auf Gruppenebene Koeffizienten $> .70$ als gut, $> .60$ als ausreichend und $< .60$ als ungenügend gelten (Evers, 2001).

2.3.2 Validität. Validität ist das wichtigste und zugleich umfassendste und komplexeste Gütekriterium: Sie bezeichnet die Übereinstimmung zwischen dem Merkmal, das gemessen werden soll, und dem tatsächlich gemessenen Merkmal und erlaubt Aussagen bezüglich der Gültigkeit und Generalisierbarkeit eines Tests (Moosbrugger & Kelava, 2012). Dabei werden fünf Aspekte unterschieden (Amelang & Schmidt-Atzert, 2006; Moosbrugger & Kelava, 2012; Schmidt-Atzert & Amelang, 2012):

(1) Augenscheinvalidität bezieht sich darauf, inwieweit der Validitätsanspruch eines Tests vom bloßen Augenschein her einem Laien gerechtfertigt erscheint. (2) Inhaltsvalidität gibt an, inwieweit ein Test das zu messende Merkmal repräsentativ erfasst. (3) Konstruktvalidität besteht, wenn vom Verhalten einer Testperson innerhalb der Testsituation auf ein zugrundeliegendes Konstrukt geschlossen werden kann: Konvergente Validität liegt vor, wenn ein Test eine hohe Übereinstimmung mit Tests für gleiche oder ähnliche Merkmale aufweist. Dagegen liegt divergente (auch diskriminante) Validität vor, wenn ein Test gegenüber Tests für andere Merkmale abgegrenzt werden kann. Weiter befasst sich faktorielle Validität mit der Faktoren- und Dimensionsstruktur eines Tests; beispielsweise kann anhand von Messinvarianzanalysen die Teststruktur bei verschiedenen Gruppen überprüft werden. (4) Die differentielle Validität ermöglicht Aussagen darüber, inwieweit ein Test zwischen verschiedenen Subgruppen zu differenzieren vermag. (5) Kriteriumsvalidität liegt schließlich vor, wenn vom Verhalten einer Testperson innerhalb der Testsituation auf ein Kriterium außerhalb der Testsituation geschlossen werden kann: Bei konkurrenter (auch Übereinstimmungs-) Validität besteht ein Zusammenhang eines Testwerts mit einem zeitgleich erfassten Kriterium respektive kann ein Testwert ein zeitgleich erfasstes Kriterium querschnittlich vorhersagen. Bei prädiktiver (auch prognostischer oder Vorhersage-) Validität hingegen besteht ein Zusammenhang eines Testwerts mit einem später erfassten Kriterium respektive kann ein Testwert ein später erfasstes Kriterium längsschnittlich vorhersagen. Zudem wird von inkrementeller—konkurrenter oder prädiktiver—Validität gesprochen, sofern ein Zuwachs an Varianzaufklärung in einem Kriterium resultiert, wenn ein Test zusätzlich zu einem anderen Test eingesetzt wird. Einschätzungen zur Augenschein- und Inhaltsvalidität werden aufgrund theoretischer Argumente vorgenommen, wohingegen Beurteilungen zur Konstrukt-, differentiellen und Kriteriumsvalidität einer

empirischen Prüfung anhand verschiedener Parameter wie Mittelwertdifferenzen, Korrelationen oder Effektstärken unterliegen (Amelang & Schmidt-Atzert, 2006; Moosbrugger & Kelava, 2012; Schmidt-Atzert & Amelang, 2012).

2.4 Validierungsbefunde der IDS

Die vorliegende Dissertation befasst sich mit der Validierung im Bereich der Konstrukt-, differentiellen und Kriteriumsvalidität der Intelligenz der IDS. Aus diesem Grund werden im Folgenden bereits bestehende Validierungsbefunde der IDS mit einem Fokus auf Intelligenz zusammengefasst, Forschungslücken aufgezeigt sowie relevante Theorien und Ergebnisse aus anderen Studien erörtert. Für die IDS-P, IDS-1 und IDS-2 liegt verschiedene Evidenz aus den Normierungs- und Validierungsstudien vor; für die IDS-1 wurden außerdem zusätzliche Studien durchgeführt.

2.4.1 Konstruktvalidität. Im Rahmen der Konstruktvalidität weisen die IDS zum einen konvergente und divergente Validität auf: Einerseits zeigen die *Kognition* und *Intelligenz* kleine bis große und die nicht-kognitiven Funktionsbereiche kleinere oder keine Zusammenhänge mit anderen Intelligenztests, andererseits finden sich für die *Intelligenz* keine oder lediglich kleine Zusammenhänge mit anderen nicht-kognitiven Tests (Grob, Gygi & Hagmann-von Arx, in Druck; Grob & Hagmann-von Arx, 2018; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx & Grob, 2014; Hagmann-von Arx, Grob, Petermann & Daseking, 2012; Hagmann-von Arx, Lemola & Grob, 2016; Hagmann-von Arx, Meyer & Grob, 2008a; Hagmann-von Arx, Petermann & Grob, 2013).

Zum anderen laden die Intelligenzuntertests aller IDS jeweils auf einen Faktor *Allgemeine Intelligenz* und die kognitiven Untertests der IDS-1 auf einen Faktor *Kognition*, was als Hinweis für faktorielle Validität verstanden werden kann (Grob & Hagmann-von Arx, 2018; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Meyer, Hagmann-von Arx, Lemola & Grob, 2010). Während für die IDS-2 die *IQ-Profil*-Struktur bestätigt werden kann und sowohl das *IQ-Profil* als auch der *IQ* Alters- und Geschlechtsinvarianz aufweisen (Grob & Hagmann-von Arx, 2018), sind für die IDS-P und IDS-1 Befunde zur Messinvarianz noch ausstehend. Ebenfalls liegt keine Evidenz vor, ob Geschlechtsdifferenzen in Testwerten bestehen.

Für ein Testverfahren ist es von großer Relevanz, sowohl die Vergleichbarkeit der Struktur (Messinvarianz) als auch potenzielle Testwertdifferenzen über verschiedene Gruppen hinweg—beispielsweise die Geschlechter—zu testen, um eine äquivalente Messung des gleichen Konstrukts unabhängig der Gruppenzugehörigkeit zu gewährleisten (Milfont & Fischer, 2010; Millsap & Kwok, 2004; Moosbrugger & Kelava, 2012). Messinvarianz stellt dabei eine Voraussetzung für die Untersuchung von Testwertdifferenzen dar, da erst bei vorhandener Invari-

anz gefolgert werden kann, dass Unterschiede in Testwerten relevant sind (American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing, 1999). So weist beispielsweise die WISC-IV nebst Altersinvarianz auch Geschlechtsinvarianz auf, sodass nachfolgende Geschlechtsdifferenzen auf Mittelwertebene als bedeutsam interpretiert werden können (Chen & Zhu, 2008; Goldbeck, Daseking, Hellwig-Brida, Waldmann & Petermann, 2010; Keith, Fine, Taub, Reynolds & Kranzler, 2006).

Zu Geschlechtsdifferenzen auf Mittelwertebene existiert eine Vielzahl an Theorien und Empirie. Gemäß der Ähnlichkeitshypothese (Hyde, 2005) sind sich die Geschlechter in der Intelligenz ähnlich, während gemäß der Differenzhypothese (Maccoby & Jacklin, 1974) Unterschiede—Stärken für Männer bei visuell-räumlichen und mathematischen Fähigkeiten sowie Stärken für Frauen bei verbalen Fähigkeiten—bestehen. Weiter weisen gemäß der Hypothese der größeren männlichen Variabilität (Hyde & Mertz, 2009) Männer eine größere Varianz in der Intelligenz auf. Schließlich besagt die Maskierungshypothese (Johnson & Bouchard, 2007), dass die allgemeine Intelligenz Geschlechtsdifferenzen in Faktoren maskieren kann, und die Entwicklungstheorie von Lynn (1994) geht davon aus, dass Geschlechtsunterschiede in Abhängigkeit des Alters variieren.

Studien im Kindes- und Jugendalter zeigen mehrheitlich, dass die allgemeine Intelligenz im Sinne der Ähnlichkeitshypothese keine, die Faktoren und Untertests im Sinne der Differenzhypothese jedoch einige kleine bis mittlere Mittelwertdifferenzen zwischen Jungen und Mädchen aufweisen: höhere Mittelwerte für Jungen in der fluiden Intelligenz und in visuell-räumlichen Fähigkeiten sowie höhere Mittelwerte für Mädchen im Kurz- und Langzeitgedächtnis und in der Verarbeitungsgeschwindigkeit (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Keith, Reynolds, Patel & Ridley, 2008; Keith, Reynolds, Roberts, Winter & Austin, 2011; Reynolds, Keith, Ridley & Patel, 2008; Savage-McGlynn, 2012; Sellers, Burns & Guyrke, 2002). Entsprechend der Hypothese der größeren männlichen Variabilität sind Varianzen in der allgemeinen Intelligenz bei Jungen grösser, während in Faktoren und Untertests keine Unterschiede bestehen (Arden & Plomin, 2006; Deary, Thorpe, Wilson, Starr & Whalley, 2003; Palejwala & Fine, 2015; Reynolds et al., 2008).

Da für die IDS-P und IDS-1 Befunde zur Messinvarianz sowie zu Geschlechtsdifferenzen in Testwerten ausstehend sind, ist ein erstes Ziel dieser Dissertation, für die Intelligenz der IDS-P und IDS-1 die Messinvarianz über das Alter und Geschlecht zu testen sowie weiterführend Geschlechtsdifferenzen auf Mittelwert- und Varianzebene zu untersuchen (Studie 1).

2.4.2. Differentielle Validität. Nebst Konstruktvalidität weisen die IDS differentielle Validität auf, da sie in der Intelligenz sowie in weiteren Funktionsbereichen zwischen verschiedenen Subgruppen zu differenzieren vermögen. Für die IDS-P liegen Vergleiche von Kindern mit Trisomie 21 (tiefere Werte im gesamten Test), mit einer allgemeinen Entwicklungsauffälligkeit (tiefere Werte im gesamten Test), mit einer Sprachentwicklungsauffälligkeit (tiefere Werte in *Sprache* und *Mathematik*) sowie von Frühgeborenen (tiefere Werte in *Kognition* und *Psychomotorik*) mit jeweils unauffälligen Kontrollkindern vor. Weiter wurden Deutsch- und Fremdsprachige verglichen, wobei letztere tiefere Werte in der *Sprache* erzielten (Grob, Reimann et al., 2013).

Für die IDS-1 gibt es Vergleiche von Kindern mit einer aggressiven Verhaltensauffälligkeit (tiefere Werte im gesamten Test), mit einer Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS; tiefere Werte in *Kognition*, *Sprache*, *Mathematik* und *Leistungsmotivation*) sowie mit Asperger-Syndrom (tiefere Werte in *Sozial-Emotionaler Kompetenz*) mit jeweils unauffälligen Kontrollkindern. Weiter weisen Kinder mit einer Hochbegabung höhere und solche mit einer Lernbehinderung tiefere Werte in der *Kognition*, *Sprache* und *Mathematik* als durchschnittlich begabte auf. Ferner wurden Deutsch- und Fremdsprachige verglichen, wobei letztere tiefere Werte in der *Sprache* und *Mathematik* erzielten (Grob, Meyer et al., 2013; Hagmann-von Arx et al., 2008a; Hagmann-von Arx et al., 2013; Meyer, Hagmann-von Arx & Grob, 2009).

Für die IDS-2 wurden Vergleiche von Kindern und Jugendlichen mit ADHS (tiefere Werte in *Intelligenz*, *Exekutiven Funktionen*, *Schulischen Kompetenzen* und *Psychomotorik*), mit einer Autismus-Spektrum-Störung (tiefere Werte in *Sozial-Emotionaler Kompetenz* und *Psychomotorik*), mit einer motorischen Auffälligkeit (tiefere Werte in *Psychomotorik*), mit einer Lese- und Rechtschreibstörung (tiefere Werte in den Untertests *Lesen* und *Rechtschreiben*) sowie von mathematisch talentierten (höhere Werte im Untertest *Logisch-Mathematisches Denken*) mit jeweils unauffälligen Kontrollgruppen durchgeführt. Weiter weisen überdurchschnittlich intelligente Kinder und Jugendliche in der *Intelligenz*, *Exekutiven Funktionen*, *Schulischen Kompetenzen* und *Psychomotorik* höhere und solche mit einer Intelligenzminderung in der *Intelligenz*, *Exekutiven Funktionen*, *Schulischen Kompetenzen*, *Psychomotorik* und *Sozial-Emotionaler Kompetenz* tiefere Werte als durchschnittlich begabte auf (Grob & Hagmann-von Arx, 2018). Zum Vergleich zwischen Deutsch- und Fremdsprachigen gibt es bisher noch keine Befunde.

Die Vergleichbarkeit von Testverfahren bei verschiedenen Sprachgruppen stellt insbesondere in der heutigen Zeit einen zentralen Validitätsaspekt dar. In den letzten Jahren ist in

Europa eine Zunahme an Personen mit Migrationshintergrund zu verzeichnen (Eurostat, 2017); für die Schweiz beispielsweise liegt deren Anteil aktuell bei einem Viertel der Bevölkerung (Bundesamt für Statistik, 2017). Folglich werden viele fremdsprachige Personen mit deutschsprachigen Tests abgeklärt. Fremdsprachige zeigen im Vergleich zu Muttersprachigen oftmals eine altersverzögerte Sprachentwicklung (Reich et al., 2002), weshalb für sie bei nicht muttersprachlichen Tests Schwierigkeiten beim Instruktionsverständnis und bei der Antwortproduktion auftreten können (Daseking & Petermann, 2015; L. G. Weiss et al., 2006), was zu einer Unterschätzung ihrer Testleistungen führen kann (Flanagan & Ortiz, 2001). Dies wiederum kann langfristig persönliche, soziale und schulische Adaptationsschwierigkeiten sowie verpasste Bildungschancen nach sich ziehen (Calero et al., 2013).

Studien zum Vergleich von muttersprachlichen und nicht muttersprachlichen Personen zeigen Differenzen in Abhängigkeit der Zusammensetzung der Testverfahren. Während die allgemeine Intelligenz der IDS-P und IDS-1 sowie des *Snijders-Oomen nonverbalen Intelligenztests 2½-7* (SON-R 2½-7; Tellegen, Laros & Petermann, 2007) und SON-R 6-40 keine Differenzen zwischen Deutsch- und Fremdsprachigen zeigt, finden sich kleine Unterschiede zuungunsten von Fremdsprachigen in der WISC-IV (Daseking, Lipsius, Petermann & Waldmann, 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013; Janke, Daseking & Petermann, 2008; Toussaint, Heinze, Lipsius & Petermann, 2012). Auf Faktor- und Untertestebene der obigen Testverfahren zeigen sich keine Differenzen im Kurzzeitgedächtnis, in der visuellen Wahrnehmung und in der Verarbeitungsgeschwindigkeit, es finden sich jedoch kleine Unterschiede zuungunsten von Fremdsprachigen in der fluiden Intelligenz, im Langzeitgedächtnis sowie in verbalen Aspekten (Daseking et al., 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013; Janke et al., 2008; Toussaint et al., 2012).

Da für die IDS-2 keine Befunde zum Vergleich zwischen Deutsch- und Fremdsprachigen vorliegen, ist ein zweites Ziel dieser Dissertation, diese Forschungslücke für die Intelligenz der IDS-2 zu schließen (Studie 2). Da zudem bisherige Studien jeweils deutsch- und fremdsprachige Personen miteinander verglichen haben, Fremdsprachigkeit jedoch teilweise unterschiedlich operationalisiert wurde, besteht ein weiteres Ziel dieser Dissertation darin, innerhalb der Fremdsprachigen zwei Substichproben zu differenzieren und diese mit Monolingualen (Personen mit Muttersprache Deutsch) zu vergleichen: Fremdsprachige (Personen mit einer anderen Muttersprache als Deutsch) und Bilinguale (Personen, die Deutsch und eine andere Sprache als Muttersprache haben; Studie 2).

2.4.3 Kriteriumsvalidität. In Bezug auf Kriteriumsvalidität zeigt die Intelligenz unter anderem Zusammenhänge mit Schulleistungen, wobei sie im Vergleich zu anderen Merkmalen als bester und stärkster Prädiktor gilt (Roth et al., 2015). Schulleistungen bezeichnen den Lern- und Schulerfolg und werden mittels Schulnoten oder standardisierter fachspezifischer Leistungstests operationalisiert (Krapp, 1997). Im Vergleich zu Leistungstests beinhalten Schulnoten eine breitere Informationsbasis und gelten als reliabler und relevanter, da sie mehr Quellen einbeziehen, Leistungen über eine längere Zeitspanne erfassen sowie ausschlaggebend für den Übertritt in höhere Klassen- und Schulstufen und für weitere akademische Qualifikationen sind (Roth et al., 2015). Eine aktuelle Metaanalyse findet einen großen Zusammenhang von Intelligenz mit der Durchschnittsnote ($p = .54$) sowie insbesondere mit den Noten in den Fächern Mathematik/Naturwissenschaften ($p = .49$) und Sprache ($p = .44$) für die Primar- und Sekundarstufe (Roth et al., 2015). Dabei hängt die Intelligenz sowohl quer- als auch längsschnittlich über ein, drei und fünf Jahre mit Schulnoten zusammen (z.B. Heaven & Ciarrochi, 2012; Heaven, Ciarrochi & Vialle, 2007; Qualter, Gardner, Pope, Hutchinson & Whiteley, 2012; Steinmayr & Spinath, 2009).

Im Rahmen der Kriteriumsvalidität für Schulleistungen liegen für die IDS-P keine Befunde vor, da die Kinder bei der Normierung noch im Vorschulalter waren und bislang keine repräsentativen Studien durchgeführt wurden. Für die IDS-2 gibt es Befunde zur konkurrenten Validität, wobei die *Intelligenz* kleine bis mittlere Zusammenhänge mit Noten sowie Elterneinschätzungen in Deutsch, Mathematik, Mensch/Umwelt, Geografie/Geschichte und Biologie/Chemie/Physik zeigt (Grob & Hagmann-von Arx, 2018). Aufgrund der erst kürzlich erfolgten Normierung der IDS-2 existieren noch keine Daten zur prädiktiven Validität.

Für die IDS-1 liegen sowohl Analysen zur konkurrenten als auch prädiktiven Validität vor. In Bezug auf die konkurrente Validität zeigen sich kleine bis große Zusammenhänge der *Kognition* mit Kompetenzwahrnehmungen von Eltern und Lehrern sowie mit Schulleistungstests für Lesen, Schreiben und Rechnen. Zudem korrelieren *Sprache* und *Mathematik* mit einem Schulleistungstest für Rechnen. Im Rahmen der prädiktiven Validität kann *Kognition* 33 bis 37% der Varianz in der Durchschnittsnote sowie in den Noten in Deutsch, Mathematik und Sachkunde drei Jahre später erklären. Zusätzlich zur *Kognition* sind *Sprache*, *Mathematik* und *Sozial-Emotionale Kompetenz* für einzelne Noten prädiktiv (Grob, Reimann et al., 2013; Gut, Reimann & Grob, 2012, 2013).

Für die IDS-1 sind allerdings Befunde dazu, ob deren prädiktive Validität vergleichbar mit derjenigen von anderen Testverfahren ist, ausstehend. Deshalb ist ein weiteres Ziel dieser Dissertation, die prädiktive Validität von anderen deutschsprachigen Intelligenztests—WISC-

IV, RIAS und SON-R 6-40—zu untersuchen und die prädiktive Validität der Intelligenz der IDS-1 derjenigen der genannten Tests gegenüberzustellen (Studie 3). Weiter stehen Befunde zur prädiktiven Validität der IDS-1 über einen Zeitraum von mehr als drei Jahren aus. Aus diesem Grund liegt ein weiteres Ziel dieser Dissertation darin, diese für die Intelligenz der IDS-1 über sieben Jahre zu untersuchen (Studie 4).

Wie aus den aufgeführten Forschungsbefunden ersichtlich wird, kann die Intelligenz als „can do“ (Gottfredson, 2002, S. 37) und „Fähigkeit, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) zwar ungefähr einen Drittel der Varianz in Schulleistungen erklären (Roth et al., 2015), dennoch bleibt ein beachtlicher Teil unaufgeklärt. Als weitere relevante Prädiktoren haben sich insbesondere die beiden Persönlichkeitsmerkmale Gewissenhaftigkeit und Motivation als „will do“ (Gottfredson, 2003, S. 369) herauskristallisiert (Richardson, Abraham & Bond, 2012; Robbins et al., 2004). Gewissenhaftigkeit ist einer der Big Five-Persönlichkeitsfaktoren (Costa & McCrae, 1992) und wird als Organisation, Zuverlässigkeit und Verantwortungsbewusstsein (Friedman & Schustack, 2004) sowie „Bereitschaft, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) definiert. Metaanalysen finden kleine bis große Zusammenhänge der Gewissenhaftigkeit mit der Durchschnittsnote ($\rho = .19 - .50$; O'Connor & Paunonen, 2007; Poropat, 2009, 2014; Richardson et al., 2012), und Einzelstudien zeigen weiter Zusammenhänge mit den Noten in Sprache, Mathematik und Naturwissenschaften (z.B. Dumfart & Neubauer, 2016; Spinath, Freudenthaler & Neubauer, 2010). Zudem erklärt Gewissenhaftigkeit zu Intelligenz inkrementelle Varianz in Schulleistungen (z.B. Di Fabio & Busoni, 2007).

Im Bereich der Motivation ist insbesondere die Leistungsmotivation für Schulleistungen von Bedeutung (Robbins et al., 2004), die als Ausdauer und Beharrlichkeit bei leistungsbezogenen Aufgaben (Wigfield & Eccles, 2000) sowie „Wille, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) beschrieben wird. Metaanalysen zeigen Zusammenhänge von Leistungsmotivation ($\rho = .30$; Robbins et al., 2004) sowie insbesondere deren Facetten Selbstwirksamkeit ($\rho = .31 - .59$; Richardson et al., 2012) und Prüfungsangst ($\rho = -.24$; Richardson et al., 2012) mit der Durchschnittsnote. Weiter finden Einzelstudien Zusammenhänge von Leistungsstreben, welche sowohl eine Facette von Gewissenhaftigkeit als auch von Leistungsmotivation darstellt (vgl. Costa & McCrae, 1992; Eccles & Wigfield, 2002), mit der Durchschnitts-, Sprach- und Mathematiknote (z.B. Steinmayr & Spinath, 2007). Ebenfalls kann Leistungsmotivation zu Intelligenz (z.B. Steinmayr & Spinath, 2009) und zu Persönlichkeitsfaktoren (z.B. Steinmayr & Spinath, 2007) inkrementelle Varianz in Schulleistungen erklären.

Für die IDS-1 liegen keine Befunde zur Kriteriumsvalidität der Intelligenz im Vergleich zu Gewissenhaftigkeit und Leistungsmotivation vor. Deshalb besteht ein abschließendes Ziel dieser Dissertation darin, diese Forschungslücke zu schließen, indem diese beiden Persönlichkeitsmerkmale als inkrementelle konkurrente Prädiktoren für Schulleistungen der Intelligenz gegenübergestellt werden (Studie 4).

3 Forschungsfragen

Die vorliegende Dissertation befasst sich mit der Konstrukt-, differentiellen und Kriteriumsvalidität der Intelligenz der IDS: Unter (I) Konstruktvalidität werden die faktorielle Validität im Sinne von Alters- und Geschlechtsinvarianz sowie Geschlechtsdifferenzen für die IDS-P und IDS-1 untersucht. Unter (II) differentieller Validität wird die Differenzierungsfähigkeit der IDS-2 hinsichtlich verschiedener Sprachgruppen überprüft. Unter (III) Kriteriumsvalidität werden schließlich zum einen (III.I) die prädiktive Validität der IDS-1 und weiterer Intelligenztests sowie zum anderen (III.II) die inkrementelle konkurrente Validität von anderen Prädiktoren (Gewissenhaftigkeit und Leistungsmotivation) für Schulleistungen analysiert. Im Folgenden werden die Forschungsfragen aufgeführt. Ein Überblick über das Dissertations-Konzept ist in Abbildung 1 ersichtlich.

I Konstruktvalidität⁵

- a. Weist die Intelligenz der IDS-P Messinvarianz über das Alter in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)
- b. Weist die Intelligenz der IDS-1 Messinvarianz über das Alter in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)
- c. Weist die Intelligenz der IDS-P Messinvarianz über das Geschlecht in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)
- d. Weist die Intelligenz der IDS-1 Messinvarianz über das Geschlecht in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)
- e. Weist die Intelligenz der IDS-P Unterschiede zwischen den Geschlechtern im Mittelwert und in der Varianz in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)
- f. Weist die Intelligenz der IDS-1 Unterschiede zwischen den Geschlechtern im Mittelwert und in der Varianz in der allgemeinen Intelligenz und in den Untertests auf? (Studie 1)

II Differentielle Validität

- a. Weist die Intelligenz der IDS-2 Unterschiede zwischen Monolingualen, Bilingualen und Fremdsprachigen im Mittelwert in der allgemeinen Intelligenz, in den Faktoren und in den Untertests auf? (Studie 2)

⁵ Die Forschungsfragen (a) bis (d) gehören zur faktoriellen Validität, während die Forschungsfragen (e) sowie (f) als Folgeanalysen der Forschungsfragen (a) bis (d) der allgemeinen Konstruktvalidität zugeordnet werden.

III Kriteriumsvalidität

III.I Prädiktive Validität

- a. Ist die Intelligenz der IDS-1 ein valider Prädiktor für längsschnittliche Schulleistungen über drei Jahre? (Studie 3)
- b. Ist die Intelligenz der IDS-1 ein valider Prädiktor für längsschnittliche Schulleistungen über sieben Jahre? (Studie 4)
- c. Sind weitere deutschsprachige Intelligenztestverfahren—WISC-IV, RIAS und SON-R 6-40—valide Prädiktoren für längsschnittliche Schulleistungen über drei Jahre? (Studie 3)
- d. Ist die prädiktive Validität der Intelligenz der IDS-1 über drei Jahre mit derjenigen von anderen deutschsprachigen Intelligenztestverfahren—WISC-IV, RIAS und SON-R 6-40—vergleichbar? (Studie 3)

III.II Konkurrente Validität

- a. Ist Gewissenhaftigkeit ein valider zu Intelligenz inkrementeller Prädiktor für konkurrente Schulleistungen? (Studie 4)
- b. Ist Leistungsmotivation ein valider zu Intelligenz und Gewissenhaftigkeit inkrementeller Prädiktor für konkurrente Schulleistungen? (Studie 4)

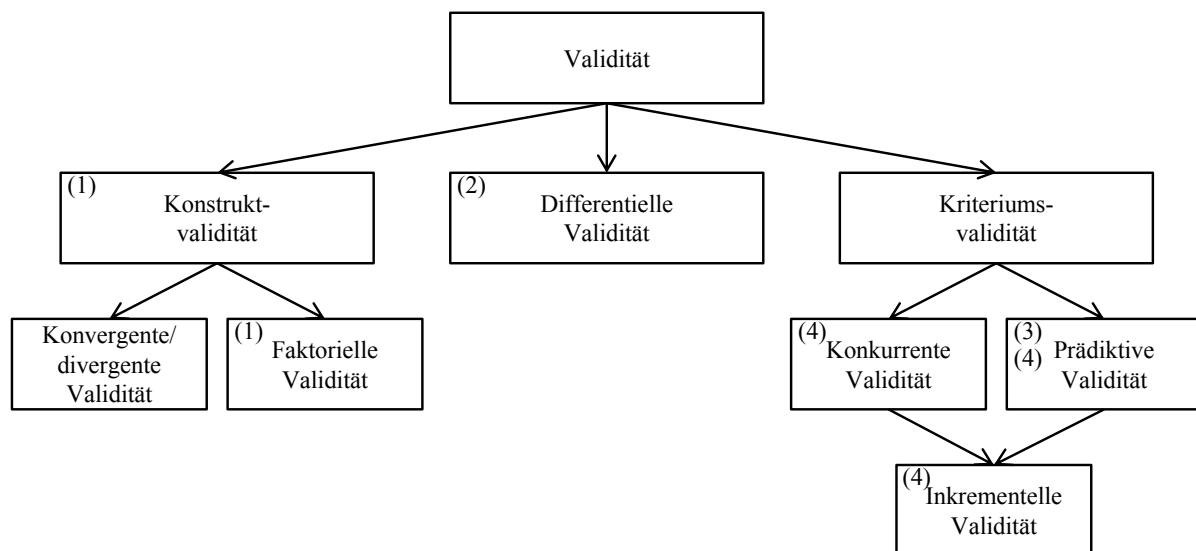


Abbildung 1. Dissertations-Konzept. Die Zahlen in Klammern repräsentieren die Studien dieser Dissertation: Studie 1 (Schweizer, Hagmann-von Arx, Lederman & Grob, 2018); Studie 2 (Schweizer, Grieder, Büniger & Grob, eingereicht); Studie 3 (Gygi, Hagmann-von Arx, Schweizer & Grob, 2017); Studie 4 (Schweizer, Gygi, Weidmann & Grob, eingereicht).

4 Methoden

4.1 Studien und Stichproben

Studie 1 (Schweizer, Hagmann-von Arx, Ledermann & Grob, 2018). Studie 1 untersuchte die faktorielle Validität im Sinne der Messinvarianz über das Alter und Geschlecht sowie Geschlechtsdifferenzen auf Mittelwert- und Varianzebene in der Intelligenz der IDS-P und IDS-1. Hierfür wurden die Normierungsstichproben der IDS-P ($N = 700$) und IDS-1 ($N = 1'330$) verwendet. Bei den IDS-P waren die Kinder zwischen 3 und 5 Jahre alt ($M = 4.46$, $SD = 0.82$). Insgesamt waren es 350 (50%) Jungen; 397 (57%) stammten aus der deutschsprachigen Schweiz, 129 (18%) aus Deutschland und 174 (25%) aus Österreich. Bei den IDS-1 waren die Kinder zwischen 5 und 10 Jahre alt ($M = 7.96$, $SD = 1.64$). Insgesamt waren es 660 (50%) Jungen; 744 (56%) stammten aus der deutschsprachigen Schweiz, 343 (26%) aus Deutschland und 243 (18%) aus Österreich. Für die Messinvarianz über das Alter wurden für die IDS-P die Altersgruppen 3;0 – 3;11 ($n = 223$), 4;0 – 4;11 ($n = 253$) und 5;0 – 5;11 ($n = 224$) und für die IDS-1 die Altersgruppen 5;0 – 6;11 ($n = 412$), 7;0 – 8;11 ($n = 492$) und 9;0 – 10;11 ($n = 426$) miteinander verglichen. Für die Messinvarianz über das Geschlecht sowie Mittelwert- und Varianzdifferenzen wurden Jungen und Mädchen miteinander verglichen.

Studie 2 (Schweizer, Grieder, Büniger & Grob, eingereicht). Studie 2 überprüfte die differentielle Validität hinsichtlich der drei Sprachgruppen Monolinguale, Bilinguale und Fremdsprachige auf Mittelwertebene in der Intelligenz der IDS-2. Hierfür wurden drei Substichproben (je $n = 143$) von Kindern und Jugendlichen der Normierungs- und Validierungsstichproben der IDS-2 ($N = 2'030$) verwendet, die nach Alter, Geschlecht und mütterlichem Ausbildungsstand gepaart wurden. In allen drei Substichproben waren die Kinder und Jugendlichen zwischen 5 und 20 Jahre alt (Monolinguale: $M = 12.45$, $SD = 4.30$; Bilinguale: $M = 12.05$, $SD = 4.32$; Fremdsprachige: $M = 12.46$, $SD = 4.32$), 70 (49%) waren Jungen, und 45 (32%) der Mütter hatten einen Hochschulabschluss. Die Probanden stammten aus der deutschsprachigen Schweiz und Deutschland. Die Monolingualen hatten Deutsch als einzige Muttersprache. Die Bilingualen sprachen gleichzeitig die Muttersprache Deutsch und eine andere Muttersprache (15% Englisch, 12% Russisch, 11% Türkisch, 10% Französisch, 10% Italienisch, 8% Spanisch und jeweils < 5% weitere Sprachen). Die Fremdsprachigen sprachen eine andere Muttersprache als Deutsch (13% Albanisch, 8% Italienisch, 8% Russisch, 8% Türkisch, 6% Bulgarisch, 6% Serbisch, 6% Tamilisch und jeweils < 5% weitere Sprachen).

Studie 3 (Gygi, Hagmann-von Arx, Schweizer & Grob, 2017). Studie 3 analysierte die prädiktive Validität der Intelligenz der IDS-1 sowie der WISC-IV, der RIAS und des SON-R 6-40 für Schulleistungen. Hierfür wurde eine Substichprobe der Normierungsstichprobe der

RIAS ($N = 2'145$) verwendet. Die Stichprobe bestand zum ersten Messzeitpunkt (2011) aus 103 Kindern zwischen 6 und 11 Jahren ($M = 9.17$, $SD = 0.93$), von welchen 54 Kinder und Jugendliche zwischen 10 und 13 Jahren ($M = 11.77$, $SD = 0.79$) zum zweiten Messzeitpunkt (2014) drei Jahre später teilnahmen. Zum zweiten Messzeitpunkt waren es 26 (48%) Jungen, und 14 (26%) der Eltern hatten einen Hochschulabschluss. Alle Probanden stammten aus der deutschsprachigen Schweiz und besuchten die Primarschule.

Studie 4 (Schweizer, Gygi, Weidmann & Grob, eingereicht). Studie 4 testete die prädiktive Validität der Intelligenz der IDS-1 sowie die inkrementelle konkurrente Validität von Gewissenhaftigkeit und Leistungsmotivation für Schulleistungen. Hierfür wurde eine Substichprobe der Normierungsstichprobe der IDS-1 ($N = 1'330$) verwendet. Die Stichprobe bestand zum ersten Messzeitpunkt (2007) aus 747 Kindern zwischen 5 und 10 Jahren ($M = 8.04$, $SD = 1.61$), von welchen 301 Kinder und Jugendliche zwischen 12 und 18 Jahren ($M = 15.45$, $SD = 1.52$) zum zweiten Messzeitpunkt (2014) sieben Jahre später teilnahmen. Zum zweiten Messzeitpunkt waren es 144 (48%) Jungen, und 69 (23%) der Mütter hatten einen Hochschulabschluss. Alle Probanden stammten aus der deutschsprachigen Schweiz und besuchten die Sekundarschule.

4.2 Messinstrumente

Studie 1. Für Studie 1 wurde die Intelligenz anhand der IDS-P (Grob, Reimann et al., 2013) für Kinder zwischen 3;0 und 5;11 Jahren sowie anhand der IDS-1 (Grob, Meyer et al., 2013) für Kinder zwischen 5;0 und 10;11 Jahren eingeschätzt. Die Intelligenz wird als allgemeine Intelligenz (*IQ*) mittels sieben Untertests erhoben (für eine detaillierte Beschreibung siehe Kapitel 2.2.4 und Tabelle 1). Die allgemeine Intelligenz ist mit IQ-Werten ($M = 100$, $SD = 15$), die Untertests sind mit Wertpunkten ($M = 10$, $SD = 3$) normiert. Die Reliabilitäten der allgemeinen Intelligenz (IDS-P: $\alpha = .91$; IDS-1: $\alpha = .92$) und der Untertests (IDS-P: $\alpha = .72 - .93$; IDS-1: $\alpha = .68 - .96$) lagen im hohen Bereich.

Studie 2. Für Studie 2 wurde die Intelligenz anhand der IDS-2 (Grob & Hagmann-von Arx, 2018) für Kinder und Jugendliche zwischen 5;0 und 20;11 Jahren eingeschätzt. Die Intelligenz wird als allgemeine Intelligenz erhoben und kann als *IQ-Profil* (14 Untertests mit 7 Faktoren), *IQ* (7 Untertests) oder *IQ-Screening* (2 Untertests) angegeben werden (für eine detaillierte Beschreibung siehe Kapitel 2.2.4 und Tabelle 2). Die allgemeine Intelligenz und die Faktoren sind mit IQ-Werten ($M = 100$, $SD = 15$) und die Untertests mit Wertpunkten ($M = 10$, $SD = 3$) normiert. Die Reliabilitäten der allgemeinen Intelligenz ($\alpha = .95 - .99$), der Faktoren ($\alpha = .92 - .98$) und der Untertests ($\alpha = .85 - .97$) lagen im hohen Bereich.

Studie 3. Für Studie 3 wurde die Intelligenz anhand der IDS-1 (Grob, Meyer et al., 2009), der WISC-IV (Petermann & Petermann, 2011), der RIAS (Hagmann-von Arx & Grob, 2014) und des SON-R 6-40 (Tellegen et al., 2012) zum ersten Messzeitpunkt und die Schulleistungen in Form von Schulnoten anhand von Elternangaben zum zweiten Messzeitpunkt eingeschätzt.

Im Rahmen der IDS-1 wurde für Studie 3 nur die allgemeine Intelligenz (*IQ*; für eine detaillierte Beschreibung siehe Studie 1, Kapitel 2.2.4 und Tabelle 1) miteinbezogen, deren Reliabilität lag im hohen Bereich ($\alpha = .92$). Die WISC-IV für Kinder und Jugendliche zwischen 6;0 und 16;11 Jahren erfasst mittels 10 bis 15 Untertests die allgemeine Intelligenz und vier Indizes, die RIAS für Personen zwischen 3;0 und 99;11 Jahren erfassen mittels vier Untertests die allgemeine Intelligenz und zwei Intelligenz-Indizes, und der SON-R 6-40 für Personen zwischen 6;0 und 40;11 Jahren erfasst mittels vier Untertests die allgemeine Intelligenz (siehe auch Kapitel 2.2.5). Für Studie 3 wurde jeweils die allgemeine Intelligenz (WISC-IV: *Gesamt-IQ*; RIAS: *Gesamtintelligenz Index*; SON-R 6-40: *SON-IQ*) berücksichtigt ($M = 100$, $SD = 15$). Die Reliabilitäten lagen im hohen Bereich (WISC-IV: $\alpha = .97$; RIAS und SON-R 6-40: $\alpha = .95$).

Die Schulleistungen wurden anhand von Elternangaben in Form von Schulnoten in den Fächern Deutsch und Mathematik erhoben. Die Eltern gaben die letzten Zeugnisnoten ihres Kindes auf einer Skala von 1 (tiefste Note) bis 6 (beste Note) an, wobei Noten zwischen 4 und 6 im genügenden Bereich liegen. Zusätzlich wurde die Durchschnittsnote der beiden Fächer berechnet.

Studie 4. Für Studie 4 wurde die Intelligenz mittels IDS-1 (Grob, Meyer et al., 2009) zum ersten Messzeitpunkt und die Gewissenhaftigkeit mittels *Fünf-Faktoren-Fragebogen für Kinder* (FFFK; Asendorpf, 1998), die Leistungsmotivation mittels *Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse* (FLM 7-13; Petermann & Winkel, 2007) und die Schulleistungen in Form von Schulnoten mittels Elternangaben zum zweiten Messzeitpunkt eingeschätzt.

Im Rahmen der IDS-1 wurde für Studie 4 nur die allgemeine Intelligenz (*IQ*; für eine detaillierte Beschreibung siehe Studie 1, Kapitel 2.2.4 und Tabelle 1) miteinbezogen, deren Reliabilität lag im hohen Bereich ($\alpha = .92$). Der FFFK ist ein Elternfragebogen und erfasst die fünf Persönlichkeitsfaktoren *Extraversion*, *Emotionale Stabilität*, *Verträglichkeit*, *Gewissenhaftigkeit* und *Kultur* mit jeweils acht Items. Dabei werden bipolare Adjektivpaare auf einer Skala von 1 (z.B. „Nachlässig“) bis 5 (z.B. „Gewissenhaft“) eingeschätzt. Für Studie 4 wurde die Gewissenhaftigkeit miteinbezogen, deren Reliabilität lag im hohen Bereich ($\alpha = .85$). Der FLM 7-13 ist ein Selbstberichtfragebogen und erfasst fünf Facetten der Leistungsmotivation

mit jeweils vier bis acht Items: *Leistungsstreben*, *Ausdauer und Fleiß*, *Angst vor Erfolg*, *Aktivierende Prüfungsangst* und *Hemmende Prüfungsangst*. Dabei werden Aussagen (z.B. „In der Schule möchte ich zu den Besten gehören“) auf einer Skala von 1 (*stimmt gar nicht*) bis 5 (*stimmt genau*) eingeschätzt. Die Reliabilitäten lagen im knapp ausreichenden bis hohen Bereich ($\alpha = .59 - .77$).

Die Schulleistungen wurden analog zu Studie 3 erhoben.

4.3 Statistische Analysen

Studie 1. In Studie 1 wurden zur Überprüfung der Messinvarianz konfirmatorische Faktoranalysen durchgeführt. Diese wurden zuerst über das Alter (Voranalysen) und anschließend über das Geschlecht (Hauptanalysen 1), jeweils zuerst auf Ebene der Untertests und anschließend auf Ebene der allgemeinen Intelligenz, ausgeführt. Zu Beginn wurden jeweils Einzelgruppenanalysen für das Ausgangsmodell (Modell 1) getestet, darauffolgend wurden je vier Modelle für Multigruppenvergleiche spezifiziert (Modell 2: Konfigurale Invarianz; Modell 3: Metrische Invarianz; Modell 4: Skalare Invarianz; Modell 5: Invarianz der Fehlervarianzen). Die Modellpassung wurde als gut respektive akzeptabel bewertet, sofern drei der vier Fit-Indizes die Grenzwerte nicht unter- respektive überschritten ($CFI \geq .95$ oder $\geq .90$, $GH \geq .95$, $NCI \geq .90$, $RMSEA \leq .06$; Hu & Bentler, 1999; Van de Schoot, Lugtig & Hox, 2012). Die Modellveränderung wurde als nicht signifikant schlechter befunden, sofern $\Delta CFI \leq .01$ war und einer der beiden anderen Fit-Indizes den Grenzwert nicht überschritt ($\Delta GH \leq .01$, $\Delta NCI \leq .02$; Cheung & Rensvold, 2002; Milfont & Fischer, 2010). Zur Untersuchung von Mittelwert- und Varianzdifferenzen (Hauptanalysen 2) wurden t -Tests und Levene-Tests durchgeführt. Zusätzlich wurden Effektstärken d nach Cohen (1988) und Variance Ratios (VRs) nach Feingold (1992) berechnet und die p -Werte der Untertests wurden mit der Korrektur nach Hommel (1988) adjustiert. Differenzen wurden als bedeutsam interpretiert, wenn sowohl Signifikanz ($p < .05$) als auch mindestens kleine Effekte ($d \geq .20$) respektive ungleiche VRs ($VR > 1.10$, $VR < 0.90$) erreicht wurden. Die Analysen erfolgten mittels AMOS (Arbuckle, 2013) und SPSS (IBM, 2012).

Studie 2. In Studie 2 wurden einfaktorielle Varianzanalysen mit Globaleffekten über alle Gruppen und anschließenden post-hoc-Kontrasten für alle Gruppenvergleiche durchgeführt. Zusätzlich wurden Effektstärken η^2 und d nach Cohen (1988) berechnet und alle p -Werte wurden mit der Korrektur nach Hommel (1988) adjustiert. Differenzen wurden als bedeutsam interpretiert, wenn sowohl Signifikanz ($p < .05$) und mindestens kleine Effekte ($\eta^2 \geq .01$) beim Globaleffekt als auch Signifikanz ($p < .05$) und mindestens kleine Effekte ($d \geq .20$) beim post-hoc-Kontrast erreicht wurden. Die Analysen erfolgten mittels R (R Core Team, 2016).

Studien 3 und 4. In Studie 3 und Studie 4 wurden hierarchische Regressionen durchgeführt. In Studie 3 wurden die Prädiktoren in separaten Modellen (Modell 1: IDS-1; Modell 2: RIAS; Modell 3: SON-R 6-40; Modell 4: WISC-IV) eingefügt (Schritt 1: Kontrollvariablen; Schritt 2: Intelligenz). In Studie 4 wurden die Prädiktoren in separaten Schritten (Schritt 1: Kontrollvariablen; Schritt 2: Intelligenz; Schritt 3: Gewissenhaftigkeit; Schritt 4: Leistungsmotivation) eingefügt. Bei Studie 3 wurde zusätzlich Bootstrapping angewendet ($N = 5'000$ zufällige Stichproben; vgl. Chernick, 2008; Efron, 1979). Effekte wurden als bedeutsam interpretiert, wenn das 95%-Konfidenzintervall des Prädiktors 0 nicht enthielt (Studie 3) respektive wenn jeweils sowohl der Regressionsmodellschritt als auch der Prädiktor Signifikanz erreichten ($p < .05$; Studie 4). Die Analysen erfolgten mittels SPSS (IBM, 2012, 2013).

5 Resultate

5.1 Konstruktvalidität

Die Resultate von Studie 1 (Schweizer et al., 2018) zeigten für die IDS-P und IDS-1 Altersinvarianz: vollständige oder partielle Invarianz der Fehlervarianzen für die Untertests sowie vollständige Invarianz der Fehlervarianzen für die allgemeine Intelligenz. Weiter wiesen die IDS-P und IDS-1 Geschlechtsinvarianz auf: vollständige oder partielle Invarianz der Fehlervarianzen für die Untertests sowie partielle Invarianz der Fehlervarianzen für die allgemeine Intelligenz. Schließlich zeigten beide Tests keine Geschlechtsdifferenzen auf Mittelwertebene in der allgemeinen Intelligenz. Auf Untertestebene fanden sich in den IDS-P höhere Mittelwerte für Jungen in *Denken Bildlich* ($d = 0.23$) sowie für Mädchen in *Aufmerksamkeit Selektiv* ($d = 0.32$) und *Gedächtnis Auditiv* ($d = 0.24$); in den IDS-1 fanden sich höhere Mittelwerte für Jungen in *Denken Bildlich* ($d = 0.45$). Die Mittelwerte der anderen Untertests sowie alle Varianzen unterschieden sich nicht bedeutsam zwischen den Geschlechtern.

5.2 Differentielle Validität

Die Resultate von Studie 2 (Schweizer, Grieder et al., eing.) zeigten, dass die Sprachgruppenzugehörigkeit 7 bis 10% der Varianz in der allgemeinen Intelligenz, zwischen 0 und 15% der Varianz in den Faktoren sowie zwischen 0 und 13% der Varianz in den Untertests erklärte, wobei die Effekte für *Denken Verbal* ($\eta^2 = .15$), *Kategorien nennen* ($\eta^2 = .12$) und *Gegenteile nennen* ($\eta^2 = .13$) am größten ausfielen. Fremdsprachige zeigten im Vergleich zu Monolingualen tiefere Mittelwerte in der allgemeinen Intelligenz ($d = 0.65 - 0.77$), im Faktor *Kurzzeitgedächtnis Auditiv* ($d = 0.49$) und in den Untertests *Zahlen- und Buchstabenreihen nachsprechen* ($d = 0.43$) und *Gemischte Zahlen- und Buchstabenreihen nachsprechen* ($d = 0.45$), im Faktor *Langzeitgedächtnis* ($d = 0.50$) und im Untertest *Geschichte nacherzählen* ($d = 0.63$), im Faktor *Denken Abstrakt* ($d = 0.45$) und im Untertest *Matrizen ergänzen* ($d = 0.42$), im Faktor *Denken Verbal* ($d = 1.01$) und in den Untertests *Kategorien nennen* ($d = 0.86$) und *Gegenteile nennen* ($d = 0.93$) sowie im Untertest *Figuren wiedererkennen* ($d = 0.41$). Weiter erreichten Fremdsprachige im Vergleich zu Bilingualen tiefere Mittelwerte in der allgemeinen Intelligenz ($d = 0.35 - 0.52$) sowie im Faktor *Denken Verbal* ($d = 0.66$) und in den Untertests *Kategorien nennen* ($d = 0.61$) und *Gegenteile nennen* ($d = 0.55$). Schließlich fanden sich zwischen Bi- und Monolingualen kleine, jedoch nicht signifikante Mittelwertdifferenzen in der allgemeinen Intelligenz zuungunsten ersterer ($d = 0.27 - 0.30$, $p > .05$). Die Mittelwerte der anderen Faktoren und Untertests unterschieden sich nicht bedeutsam zwischen den drei Sprachgruppen.

5.3 Kriteriumsvalidität: Prädiktive Validität

Die Resultate von Studie 3 (Gygi et al., 2017) und Studie 4 (Schweizer, Gygi et al., eing.) ergaben prädiktive Validität für die Intelligenz der IDS-1 für die Durchschnitts- ($\beta = .18$, $R^2 = 14\%$), Deutsch- ($\beta = .32$, $R^2 = 9\%$) und Mathematiknote ($\beta = .38$, $R^2 = 13\%$) über drei Jahre (Studie 3) sowie ebenfalls für die Durchschnitts- ($\beta = .26$, $R^2 = 7\%$), Deutsch- ($\beta = .22$, $R^2 = 5\%$) und Mathematiknote ($\beta = .22$, $R^2 = 5\%$) über sieben Jahre (Studie 4). Weiter zeigte Studie 3 prädiktive Validität für die WISC-IV ($\beta = .13$, $R^2 = 7\%$), die RIAS ($\beta = .17$, $R^2 = 10\%$) und den SON-R 6-40 ($\beta = .13$, $R^2 = 8\%$) für die Durchschnittsnote über drei Jahre. Die RIAS waren zudem für die Deutsch- ($\beta = .38$, $R^2 = 10\%$) und Mathematiknote ($\beta = .28$, $R^2 = 6\%$) und der SON-R 6-40 für die Mathematik- ($\beta = .28$, $R^2 = 8\%$), nicht aber Deutschnote, prädiktiv. Die WISC-IV konnte keine Varianz in der Deutsch- und Mathematiknote erklären.

5.4 Kriteriumsvalidität: Konkurrente Validität

Die Resultate von Studie 4 (Schweizer, Gygi et al., eing.) ergaben inkrementelle konkurrente Validität für Gewissenhaftigkeit für die Durchschnitts- ($\beta = .31$, $\Delta R^2 = 9\%$), Deutsch- ($\beta = .24$, $\Delta R^2 = 5\%$) und Mathematiknote ($\beta = .28$, $\Delta R^2 = 7\%$) über Intelligenz hinaus. Ebenfalls wies Leistungsmotivation inkrementelle konkurrente Validität für die Durchschnitts- ($\Delta R^2 = 8\%$), Deutsch- ($\Delta R^2 = 5\%$) und Mathematiknote ($\Delta R^2 = 7\%$) über Intelligenz und Gewissenhaftigkeit hinaus auf: Leistungsstreben erwies sich als Prädiktor für die Durchschnitts- ($\beta = .26$), Deutsch- ($\beta = .21$) und Mathematiknote ($\beta = .22$) und hemmende Prüfungsangst als Prädiktor für die Durchschnitts- ($\beta = -.15$) und Mathematik- ($\beta = -.19$), nicht aber Deutschnote. Die anderen drei Facetten erklärten keine Varianz in den Noten.

6 Diskussion

Das Ziel der vorliegenden Dissertation bestand darin, den Forschungsstand zur Konstrukt-, differentiellen und Kriteriumsvalidität der Intelligenz der IDS zu erweitern. Hierfür wurde erstens die Konstruktvalidität der IDS-P und IDS-1 untersucht. Zweitens wurde die differentielle Validität der IDS-2 hinsichtlich verschiedener Sprachgruppen überprüft. Drittens wurde im Rahmen der Kriteriumsvalidität die prädiktive Validität der IDS-1 für Schulleistungen über drei und über sieben Jahre analysiert. Weiter wurde die prädiktive Validität der IDS-1 über drei Jahre derjenigen der WISC-IV, der RIAS und des SON-R 6-40 gegenübergestellt. Schließlich wurde die zu Intelligenz inkrementelle konkurrente Validität von Gewissenhaftigkeit und Leistungsmotivation getestet.

6.1 Konstruktvalidität

Im Rahmen der Konstruktvalidität (Studie 1) wurde die faktorielle Validität der IDS-P und IDS-1 bestätigt, da die Intelligenz jeweils Alters- und Geschlechtsinvarianz aufwies. Dies bedeutet, dass die Teststruktur über das Alter und Geschlecht vergleichbar ist und dass das gleiche Konstrukt äquivalent gemessen wird (Millsap & Kwok, 2004). Weiter ist dadurch die Voraussetzung für die Untersuchung von Geschlechtsdifferenzen auf Mittelwert- und Varianzebene gegeben (vgl. American Educational Research Association et al., 1999).

Hinsichtlich Geschlechtsdifferenzen auf Mittelwert- und Varianzebene zeigten beide Tests keine Mittelwertdifferenzen in der allgemeinen Intelligenz, was mit der Ähnlichkeitshypothese (Hyde, 2005) und früheren Studien einhergeht (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Savage-McGlynn, 2012; Sellers et al., 2002). Ebenfalls fanden sich keine Mittelwertunterschiede in den Untertests *Wahrnehmung Visuell*, *Gedächtnis Phonologisch*, *Gedächtnis Räumlich-Visuell* und *Denken Konzeptuell*. Dies ist mit einigen anderen Befunden konform (z.B. Camarata & Woodcock, 2006; Keith et al., 2011; Sellers et al., 2002), wobei es auch Studien gibt, die in diesen Aspekten Unterschiede gefunden haben (z.B. Keith et al., 2008; Reynolds et al., 2008). Mögliche Gründe für die unterschiedlichen Ergebnisse könnten beispielsweise in der Stichprobenzusammensetzung, in verschiedenen Messmethoden oder in unterschiedlichen Analysemethoden liegen.

In beiden Tests zeigten sich kleine Mittelwertdifferenzen zugunsten von Jungen in *Denken Bildlich* sowie in den IDS-P kleine Mittelwertdifferenzen zugunsten von Mädchen in *Aufmerksamkeit Selektiv* und *Gedächtnis Auditiv*. Das Resultat zu *Denken Bildlich* deckt sich mit der Differenzhypothese (Maccoby & Jacklin, 1974) und Befunden, dass Jungen höhere Werte als Mädchen in visuell-räumlichen Fähigkeiten aufweisen (Reynolds et al., 2008). Eine mögliche Erklärung für Differenzen in *Denken Bildlich*—nicht aber in *Wahrnehmung Visuell* und

Gedächtnis Räumlich-Visuell—ist, dass *Denken Bildlich* komplexere visuell-räumliche Fähigkeiten wie räumliches Vorstellungsvermögen und mentale Rotation erfasst, während die anderen beiden Untertests eher basale Aufgaben beinhalten. Dies würde Studien entsprechen, die insbesondere bei komplexen visuell-räumlichen Aufgaben Unterschiede oder zumindest Tendenzen zugunsten von Jungen gefunden haben (Keith et al., 2008; Keith et al., 2011; Reynolds et al., 2008). Die Ergebnisse zu *Aufmerksamkeit Selektiv* und *Gedächtnis Auditiv* der IDS-P sind mit Befunden konsistent, dass Mädchen höhere Werte als Jungen in der Verarbeitungsgeschwindigkeit und im Langzeitgedächtnis aufweisen (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Keith et al., 2008; Keith et al., 2011). Ein möglicher Grund für Differenzen in den IDS-P—nicht aber in den IDS-1—könnte sein, dass Unterschiede in diesen Aufgaben altersspezifisch sind. Dies würde mit der Entwicklungstheorie von Lynn (1994) sowie Studien, in welchen Geschlechtsunterschiede in Abhängigkeit des Alters variieren, übereinstimmen (z.B. Keith et al., 2011). Da die vorliegende Dissertation keine Altersgruppen untersucht hat, müsste diese Hypothese in weiteren Untersuchungen überprüft werden.

Schließlich zeigten beide Tests in der allgemeinen Intelligenz und in allen Untertests vergleichbare Varianzen für die Geschlechter. Ähnliche Varianzen in der allgemeinen Intelligenz widersprechen der Hypothese der größeren männlichen Variabilität (Hyde & Mertz, 2009) und anderen Studien (Arden & Plomin, 2006; Deary et al., 2003), während ähnliche Varianzen in den Untertests mit anderen Studien übereinstimmen (Palejwala & Fine, 2015; Reynolds et al., 2008). Die hier gefundenen Resultate können möglicherweise durch die Stichprobensammensetzung begründet werden: Die Rekrutierung erfolgte ohne expliziten Einbezug von Sonderschulen, weshalb Kinder mit einer Intelligenzminderung oder Hochbegabung möglicherweise unterrepräsentiert sind (Varianzeinschränkung).

6.2 Differentielle Validität

Im Kontext der differentiellen Validität (Studie 2) erklärte die Sprachgruppenzugehörigkeit 7 bis 10% der Varianz in der allgemeinen Intelligenz: Es zeigten sich tiefere Werte für Fremdsprachige im Vergleich zu Mono- und Bilingualen mit kleinen bis mittleren Effekten sowie tiefere Werte für Bi- im Vergleich zu Monolingualen mit kleinen, jedoch nicht signifikanten Effekten. Diese Befunde sind mit der WISC-IV vergleichbar, während die IDS-P und IDS-1 sowie der SON-R 2½-7 und SON-R 6-40 keine Unterschiede zwischen Deutsch- und Fremdsprachigen aufwiesen (Daseking et al., 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013; Janke et al., 2008; Toussaint et al., 2012)⁶. Die

⁶ Die zitierten Studien haben jeweils Deutsch- und Fremdsprachige, letztere in der Regel definiert als Personen mit Migrationshintergrund mit Zweitsprache Deutsch, miteinander verglichen. Insofern ist ein Vergleich zwischen der Studie dieser Dissertation und den anderen Studien nicht abschließend möglich.

IDS-2 erfassen ähnlich der WISC-IV sowohl fluide als auch kristalline Intelligenzaspekte und weisen somit eine höhere Sprachgebundenheit als die IDS-P und IDS-1 sowie der SON-R 2½-7 und SON-R 6-40 auf. Folglich ist es naheliegend, dass in der vorliegenden Studie monolingual Deutschsprachige in der allgemeinen Intelligenz die höchsten Werte erzielten, gefolgt von Zwei- und zuletzt Fremdsprachigen.

Die Faktoren *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* zeigten keine Gruppendifferenzen. Dies geht mit Befunden zur visuellen Wahrnehmung und Verarbeitungsgeschwindigkeit sowie den IDS-P und IDS-1 einher (Daseking et al., 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013; Janke et al., 2008) und ist dadurch erklärbar, dass diese Faktoren relativ sprachunabhängig sind (McGrew, 2005). Ebenfalls fanden sich keine Unterschiede im Faktor *Kurzzeitgedächtnis Räumlich-Visuell* und in dessen zweitem Untertest *Rotierte Figuren wiedererkennen*; der erste Untertest *Figuren wiedererkennen* zeigte dagegen kleine Differenzen zuungunsten von Fremdsprachigen im Vergleich zu Monolingualen. Diese Resultate sind teilweise mit den IDS-P und IDS-1 vergleichbar (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). Ein möglicher Grund für Unterschiede in *Figuren wiedererkennen* ist, dass die Instruktion sprachlich sehr komplex ist. Eine Erklärung für fehlende Unterschiede in *Rotierte Figuren wiedererkennen* könnte dagegen sein, dass ein Lerneffekt vom ersten (*Figuren wiedererkennen*) zum zweiten (*Rotierte Figuren wiedererkennen*) Untertest stattgefunden hat.

Der Faktor *Kurzzeitgedächtnis Auditiv* und beide zugehörigen Untertests, der Faktor *Langzeitgedächtnis* und der Untertest *Geschichte nacherzählen* (nicht aber *Bild beschreiben*) sowie der Faktor *Denken Abstrakt* und der Untertest *Matrizen ergänzen* (nicht aber *Unpassende Bilder ausschließen*) zeigten kleine bis mittlere Differenzen zuungunsten von Fremdsprachigen im Vergleich zu Monolingualen. Das Ergebnis zu *Kurzzeitgedächtnis Auditiv* steht im Widerspruch zu den IDS-P und IDS-1 sowie zur WISC-IV (Daseking et al., 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013). Dies lässt sich möglicherweise damit erklären, dass die IDS-2 sowohl Zahlen als auch Buchstaben, sowohl einfache als auch gemischte Reihen und sowohl Vorwärts- als auch Rückwärtsnachsprechen erfassen, was womöglich höhere Sprachkompetenzen als bei den IDS-P und IDS-1 erfordert. Obwohl in der WISC-IV ähnliche Aufgaben wie in den IDS-2 gestellt werden, traten dort keine Unterschiede zwischen Deutsch- und Fremdsprachigen auf (Daseking et al., 2008). Das Resultat zu *Langzeitgedächtnis* ist mit den IDS-P und IDS-1 vergleichbar (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). *Geschichte nacherzählen* erfasst phonologische Gedächtnisaspekte mittels sprachlicher Darbietung und sprachlichem Abruf—im Gegensatz dazu erhebt *Bild beschreiben*

visuell-räumliche Aspekte mittels visueller Darbietung und visuellem sowie sprachlichem Abruf. Insofern setzt *Geschichte nacherzählen* im Vergleich zu *Bild beschreiben* höhere Sprachkenntnisse voraus. *Denken Abstrakt* schließlich ist mit dem WISC-IV-Index *Wahrnehmungsgeladenes Logisches Denken* vergleichbar, wobei sich in den Matrizen-Untertests der WISC-IV sowie des SON-R 6-40 im Gegensatz zu den IDS-2 keine Unterschiede zeigten (Daseking et al., 2008; Toussaint et al., 2012). Unterschiede im ersten Untertest *Matrizen ergänzen* könnten dadurch begründet werden, dass auch bei sprachfreien Tests sprachliche Leistungen erforderlich sind, da Testsprache und -prozess verstanden werden müssen (Gunderson & Siegel, 2001). Weiter sind viele Aufgaben von aufgabenspezifischen Vorerfahrungen abhängig (Griorenko & Sternberg, 1998); möglicherweise sind die IDS-2-Matrizen für Fremdsprachige ungewohnt. Dahingegen könnten fehlende Effekte im zweiten Untertest *Unpassende Bilder ausschließen* darauf zurückzuführen sein, dass—ähnlich zu *Kurzzeitgedächtnis Räumlich-Visuell*—ein Lerneffekt stattgefunden hat.

Schließlich resultierten im Faktor *Denken Verbal* sowie in den Untertests *Kategorien nennen* und *Gegenteile nennen* mit 12 bis 15% erklärter Varianz durch die Sprachgruppenzugehörigkeit die größten Gruppendifferenzen mit mittleren bis großen Unterschieden zuungunsten von Fremdsprachigen im Vergleich zu Mono- sowie Bilingualen. Diese Ergebnisse sind in Übereinstimmung mit der WISC-IV (Daseking et al., 2008) und damit erklärbar, dass in diesen IDS-2-Skalen kristalline und sprachgebundene Intelligenzaspekte erhoben werden.

6.3 Kriteriumsvalidität: Prädiktive Validität

Im Rahmen der Kriteriumsvalidität wies die Intelligenz der IDS-1 prädiktive Validität für die Durchschnitts-, Deutsch- und Mathematiknote über drei (Studie 3) und über sieben Jahre (Studie 4) auf. Dies entspricht der Literatur zur Intelligenz als längsschnittlichem Prädiktor für Schulleistungen (z.B. Heaven & Ciarrochi, 2012; Heaven et al., 2007; Qualter et al., 2012). Intelligenz begünstigt die Auffassungsgabe und schulische Lernfähigkeit (vgl. Di Fabio & Busoni, 2007) und weist eine relativ hohe Stabilität auf (Rost, 2009). Folglich ist es plausibel, dass sich längsschnittliche Zusammenhänge mit Schulleistungen zeigen. Die Effekte beider Studien waren klein (Studie 4) respektive klein bis mittel (Studie 3). Sie sind somit vergleichbar mit denjenigen der anderen Studien der IDS-1 (Gut et al., 2012, 2013), jedoch kleiner als denjenigen der Metaanalyse von Roth et al. (2015). Die eher kleinen Effekte könnten dadurch begründet sein, dass in Studie 3 die Varianz der Schulleistungen eingeschränkt war, da alle angegebenen Noten im genügenden Bereich lagen, und dass in Studie 4 das Zeitintervall mit sieben Jahren relativ groß war. Sowohl eine Varianzeinschränkung als auch vergrößerte Zeitintervalle können mit einer Verringerung von Zusammenhängen einhergehen (Howitt & Cramer, 2008).

Zusätzlich zu den IDS-1 wiesen die WISC-IV, die RIAS und der SON-R 6-40 prädiktive Validität für die Durchschnittsnote über drei Jahre auf. Die RIAS waren zudem für die Deutsch- und Mathematiknote und der SON-R 6-40 für die Mathematiknote prädiktiv (Studie 3). Dies geht mit der Literatur einher, in der verschiedene Intelligenztests sowie verschiedene Testtypen (verbal, nonverbal oder gemischt) prädiktiv für Schulleistungen sind (vgl. Roth et al., 2015). Die Effekte für die WISC-IV, die RIAS und den SON-R 6-40 waren klein bis mittel und somit im Bereich derjenigen der IDS-1. Die unterschiedlichen Effekte der Tests für die verschiedenen Schulnoten liegen womöglich an deren Zusammensetzung: Die Literatur zeigt, dass insbesondere das Kurzzeit- und Arbeitsgedächtnis Zusammenhänge mit schulischem Lernen zeigen— dabei ist das phonologische und verbale Gedächtnis mit der Sprache, das visuell-räumliche Gedächtnis hingegen stärker mit der Mathematik assoziiert (Dehn, 2008). Der IDS-1-*IQ* besteht aus einem vergleichsweise großen Gedächtnisanteil und dieser wird phonologisch sowie visuell-räumlich erfasst, was die Zusammenhänge mit der Deutsch- und Mathematiknote sowie dem Durchschnitt der beiden Noten erklärt. Der *Gesamtintelligenz Index* der RIAS beinhaltet verbale Fähigkeiten und sprachlogisches Denken sowie visuell-räumliche Fähigkeiten, sodass es ebenfalls naheliegend ist, dass sich Zusammenhänge mit der Deutsch- und Mathematik- sowie der Durchschnittsnote zeigen. Im Gegensatz dazu setzt sich der *SON-IQ* aus nonverbalen und visuell-räumlichen Aufgaben zusammen, allerdings ohne Gedächtnisaspekte, was womöglich zu Effekten für die Mathematiknote, aber fehlenden Effekten für die Deutschnote geführt hat. Schließlich schätzt die WISC-IV das phonologische, nicht aber das visuell-räumliche Gedächtnis ein, was möglicherweise die fehlenden Effekte für die Mathematiknote erklären kann. Obwohl sich deren *Gesamt-IQ* aus zwei sprachlichen Indizes—(phonologisches) *Arbeitsgedächtnis* und *Sprachverständnis*—zusammensetzt, waren diese in der vorliegenden Studie nicht ausreichend für eine Varianzaufklärung in der Deutschnote. Die Stichprobe von Studie 3 war mit $N = 54$ relativ klein und die Teststärke entsprechend zu gering, um kleine Effekte zu entdecken. Die vorliegenden Befunde müssten somit in einer weiteren Studie mit einer größeren Stichprobe repliziert werden.

6.4 Kriteriumsvalidität: Konkurrente Validität

Im Kontext der Kriteriumsvalidität zeigte Gewissenhaftigkeit inkrementelle konkurrente Validität für die Durchschnitts-, Deutsch- und Mathematiknote über Intelligenz hinaus (Studie 4). Dies ist konform mit Befunden, dass Gewissenhaftigkeit ein querschnittlicher Prädiktor für Schulleistungen ist (z.B. Poropat, 2009, 2014) und zu Intelligenz inkrementelle Varianz erklären kann (z.B. Di Fabio & Busoni, 2007). Mögliche Gründe für den Zusammenhang zwischen Gewissenhaftigkeit und Schulleistungen sind, dass eine hohe Gewissenhaftigkeit mit

mehr Organisation, Sorgfalt und Ausdauer im Schulkontext (z.B. in Schulstunden oder bei Hausaufgaben) einhergeht oder Lehrerbeurteilungen beeinflussen kann (vgl. Di Fabio & Busoni, 2007; Rosander & Bäckström, 2014). Die Effekte für Gewissenhaftigkeit waren klein bis mittel und die erklärte Varianz lag im Bereich derjenigen der Intelligenz. Da in der vorliegenden Studie Gewissenhaftigkeit quer-, Intelligenz hingegen längsschnittlich erfasst wurde, sind die Effektstärken der beiden Prädiktoren allerdings nicht direkt vergleichbar.

Weiter zeigten Leistungsmotivation-Facetten inkrementelle konkurrente Validität für Schulleistungen über Intelligenz und Gewissenhaftigkeit hinaus: Leistungsstreben war ein Prädiktor für die Durchschnitts-, Deutsch- und Mathematiknote und hemmende Prüfungsangst ein Prädiktor für die Durchschnitts- und Mathematiknote (Studie 4). Dies ist konsistent mit Befunden, dass Leistungsmotivation einen querschnittlichen Prädiktor für Schulleistungen darstellt (z.B. Robbins et al., 2004) und zu Intelligenz sowie zu Persönlichkeitsfaktoren inkrementelle Varianz erklären kann (z.B. Steinmayr & Spinath, 2007, 2009). Weiter ist dies konform mit Studien, die aufzeigen, dass insbesondere die Facetten Leistungsstreben und Prüfungsangst Zusammenhänge mit Schulleistungen zeigen (z.B. Richardson et al., 2012; Steinmayr & Spinath, 2007, 2009). Der Zusammenhang zwischen Leistungsstreben und Schulleistungen kann zum einen dadurch erklärt werden, dass Leistungsstreben ebenfalls eine Facette von Gewissenhaftigkeit darstellt (Costa & McCrae, 1992), welche mit Schulleistungen assoziiert ist (z.B. Poropat, 2009). In der vorliegenden Studie korrelierten die beiden Variablen hingegen nicht signifikant miteinander, was sich allerdings durch die verschiedenen Informationsquellen begründen lässt. Zum anderen ist es denkbar, dass Schüler mit hohem Leistungsstreben mehr Zeit für die Schule investieren (z.B. bei Hausaufgaben oder der Testvorbereitung), da sie mehr als andere erreichen wollen (vgl. Petermann & Winkel, 2007). Im Gegensatz dazu lassen sich die Effekte von hemmender Prüfungsangst möglicherweise durch Leistungsangst erklären: Sprachliche Fächer erfordern in der Regel eine schlüssige Argumentation, mathematische Fächer hingegen eher spezifische Antworten, die entweder richtig oder falsch sind. Womöglich trauen sich Schüler mit Leistungsangst nicht, Antworten in Mathematik zu geben, und somit kommt Prüfungsangst eine größere Bedeutung in Mathematik als in Deutsch zu (vgl. Steinmayr & Spinath, 2007). Insgesamt waren die Effekte für Leistungsmotivation klein und die erklärte Varianz lag im Bereich derjenigen der Gewissenhaftigkeit und Intelligenz.

6.5 Implikationen

Basierend auf den Befunden der vorliegenden Dissertation lassen sich verschiedene Implikationen für die Praxis ableiten. Für die IDS-P und IDS-1 wurde die Konstruktvalidität bestätigt, da die Intelligenz Alters- und Geschlechtsinvarianz aufweist sowie lediglich wenige und

maximal kleine Geschlechtsdifferenzen auf Mittelwertebene bestehen. Dies impliziert, dass die beiden Tests für alle postulierten Altersgruppen und beide Geschlechter verwendet werden können sowie für beide Geschlechter mittels geschlechtsunabhängiger Normtabellen eine faire Intelligenzeinschätzung möglich ist. Nichtsdestotrotz sollen diejenigen Untertests, welche Differenzen zu(un)gunsten eines Geschlechts aufweisen—*Aufmerksamkeit Selektiv*, *Gedächtnis Auditiv* und *Denken Bildlich*—, mit einer gewissen Vorsicht interpretiert werden.

Für die IDS-2 zeigte der Sprachgruppenvergleich im Rahmen der differentiellen Validität, dass die Gruppenzugehörigkeit je nach Faktor und Untertest zwischen 0 und 15% der Varianz erklärt: Während *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* nicht respektive nur wenig sprachgebunden sind, weisen *Kurzzeitgedächtnis Auditiv*, *Kurzzeitgedächtnis Räumlich-Visuell*, *Langzeitgedächtnis*, *Denken Abstrakt* und *Denken Verbal* eine höhere Sprachgebundenheit sowie Aufgaben- und Instruktionskomplexität auf. Dabei sind bei bestimmten Faktoren beide zugehörigen Untertests ähnlich sprachungebunden (z.B. *Verarbeitung Visuell*) respektive sprachgebunden (z.B. *Denken Verbal*), wohingegen bei anderen Faktoren Unterschiede zwischen den Untertests hinsichtlich der Sprachgebundenheit bestehen (z.B. *Langzeitgedächtnis*). Dies impliziert, dass der sprachliche Hintergrund einer Person und damit zusammenhängend mögliche Nachteile für Personen mit nicht (ausschließlich) deutscher Muttersprache bei der Testwertinterpretation berücksichtigt werden sollen. Sofern (große) intraindividuelle Leistungsdiskrepanzen vorliegen, können diese womöglich auf die unterschiedliche Sprachgebundenheit der Faktoren und zugehörigen Untertests zurückgeführt werden, und die allgemeine Intelligenz sollte aufgrund ihres kristallinen Anteils mit Vorsicht interpretiert werden.

Die IDS-2 gehören zu sogenannten verbalen Tests (verbale Instruktionen und/oder Aufgabenstellungen und/oder Antworten; vgl. Tellegen et al., 2012) sowie Statustests (kein Feedback; vgl. Guthke, Beckmann & Wiedl, 2003). Bei solchen besteht an sich die Problematik, „dass bei Personen, deren Muttersprache nicht Testsprache ist, unwillkürlich auch die mangelhafte produktive bzw. rezeptive Sprachfähigkeit in die Testwerte einfließt“ (Moosbrugger & Höfling, 2006, S. 409). Mögliche Alternativen stellen nonverbale und kulturfaire sowie dynamische Tests respektive Lerntests dar: Nonverbale und kulturfaire Tests wie beispielsweise der SON-R 6-40 (Tellegen et al., 2012) respektive der *Culture Fair Test – Skala 2* (CFT 20-R; R. H. Weiss, 2006) minimieren durch individuell auf das Sprachvermögen des Probanden angepasste Instruktionen (verbal, nonverbal oder kombiniert verbal-nonverbal) sowie mehrheitlich sprachfreie respektive kulturfaire Aufgabenstellungen und sprachfreie Antworten eine Benachteiligung von Personen mit Migrationshintergrund. Im Vergleich dazu bauen dynamische Tests

Denkhilfen oder Feedbacks in den Testprozess ein, um das Lernpotenzial sowie die maximal mögliche Performanz zu erfassen und um eine validere Intelligenzeinschätzung, insbesondere bei Minoritäten, zu erreichen (Guthke & Wiedl, 1996). Für die IDS-2 ist an der Universität Basel eine Adaptation in Form der *Migrationsfairen IDS-2* in Vorbereitung, indem für Personen mit inexistenten oder minimalen Deutschkenntnissen für fünf Intelligenzuntertests—sowie zusätzlich zwei *Exekutive Funktionen*-Untertests—dynamische Untertestversionen entwickelt werden. Obwohl das dynamische Testen einen vielversprechenden Ansatz darstellt und Studien zeigen, dass insbesondere Probanden mit Migrationshintergrund von Lerntests profitieren (z.B. Hessels, 1997), existieren bislang nur wenige geeignete Verfahren dazu (Börnert & Wilbert, 2016). Ein Ziel für die Zukunft sollte deshalb sein, weitere dynamische Tests zu entwickeln.

Für die IDS-1 wurde schließlich im Rahmen der Kriteriumsvalidität die prädiktive Validität nachgewiesen, indem die Intelligenz Schulleistungen in Form von Noten (Durchschnitt, Deutsch und Mathematik) sowohl über drei als auch über sieben Jahre vorhersagen kann. Zusätzlich zeigten die WISC-IV, die RIAS und der SON-R 6-40 prädiktive Validität für die Durchschnittsnote sowie, je nach Test, für die Deutsch- und Mathematiknote. Die Effektstärken lagen jeweils im ähnlichen Bereich. Dies impliziert, dass alle Intelligenztests geeignet sind, um mittel- bis langfristige Vorhersagen bezüglich Schulleistungen zu machen. Weiter impliziert dies, dass die prädiktive Validität der IDS-1 mit derjenigen der anderen Tests hinsichtlich der Effektstärken vergleichbar ist. Die unterschiedliche Zusammensetzung der Tests (u.a. in Bezug auf das Kurzzeit- und Arbeitsgedächtnis) sowie die damit zusammenhängenden unterschiedlichen Zusammenhänge der Tests für spezifische Noten (IDS-1 und RIAS: Durchschnitt, Deutsch und Mathematik; SON-R 6-40: Durchschnitt und Mathematik; WISC-IV: Durchschnitt) sollten jedoch berücksichtigt werden.

Zusätzlich zur Intelligenz der IDS-1 wiesen Gewissenhaftigkeit sowie einzelne Leistungsmotivation-Facetten inkrementelle konkurrente Validität für Noten auf. Dies wiederum bedeutet, dass die Intelligenz als „can do“ (Gottfredson, 2002, S. 37) und „Fähigkeit, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) zwar einen bedeutenden Varianzanteil in Schulleistungen erklären kann, dass aber gleichzeitig andere Konstrukte wie Gewissenhaftigkeit als „Bereitschaft, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) sowie Leistungsmotivation als „Wille, akademische Leistungen zu erbringen“ (Di Domenico & Fournier, 2015, S. 156) ebenfalls relevant sind. Entsprechend sollen in der Praxis bei Fragestellungen sowie bei möglichen Fördermaßnahmen hinsichtlich Schulleistungen zusätzlich zur Intelligenz weitere Aspekte berücksichtigt und integriert wer-

den, um ein möglichst umfassendes Bild zu erhalten (vgl. auch Grob, Hagmann-von Arx et al., 2009).

Es ist abschließend anzumerken, dass die in dieser Dissertation gezogenen Schlussfolgerungen teilweise anders hätten ausfallen können, je nach Konzeptualisierung der Validität (vgl. auch Kane, 2001): Die Untersuchung von Geschlechtsdifferenzen als Folgeanalysen der Messinvarianz wurde der Konstruktvalidität zugeordnet und basierend auf den Ergebnissen wird geschlussfolgert, dass die IDS-P und IDS-1 konstruktvalide sind. Eine alternative Zuordnung wäre diejenige zur differentiellen Validität oder zum Nebengütekriterium Testfairness (keine Benachteiligung von Personen aufgrund ihrer Zugehörigkeit zu spezifischen Gruppen; vgl. Amelang & Schmidt-Atzert, 2006; Moosbrugger & Kelava, 2012). Je nach Interpretation würden die IDS-P und IDS-1 als differentiell valide—da sie in einigen Untertests zwischen den Geschlechtern zu differenzieren vermögen—, jedoch nicht testfair—da in einigen Untertests (kleine) Differenzen zwischen den Geschlechtern auftreten—gelten. Weiter wurde die Analyse von Sprachgruppendifferenzen der differentiellen Validität zugeordnet. Es stellt sich jedoch die Frage, ob eine Differenzierung zwischen Gruppen die Validität stützt (Unterscheidung zwischen verschiedenen Sprachgruppen möglich)—oder aber, ob auftretende Unterschiede die Validität und somit Generalisierbarkeit hindern (Nachteile in einzelnen Untertests und Faktoren in Abhängigkeit der Sprachgruppenzugehörigkeit). Somit kann für die vorliegende Dissertation nicht abschließend beantwortet werden, ob die IDS-2 im Hinblick auf Sprachgruppen differentiell valide sind oder nicht—je nach Interpretation könnten die vorliegenden Ergebnisse als Validitätsnachweis oder aber -einschränkung angesehen werden. Alternativ könnte die Frage nach Sprachgruppendifferenzen auch dem Nebengütekriterium Testfairness zugeordnet werden (vgl. Moosbrugger & Kelava, 2012), wobei die IDS-2 wohl als nicht vollständig testfair gelten würden, da in einigen Untertests und Faktoren (kleine bis große) Differenzen zwischen den Sprachgruppen auftreten.

6.6 Stärken und Limitationen

Die vorliegende Dissertation weist verschiedene Stärken und Limitationen auf. Eine erste Stärke ist der Einbezug aller deutschsprachigen IDS-Versionen: Es wurden sowohl die IDS-P (Studie 1) als auch die IDS-1 (Studie 1, 3 und 4) sowie die IDS-2 (Studie 2) miteinbezogen und weiterführend validiert. Eine zweite Stärke betrifft die Untersuchung verschiedener Validitätsaspekte: Es wurde sowohl die Konstruktvalidität (Alters- und Geschlechtsinvarianz sowie Geschlechtsdifferenzen; Studie 1) als auch die differentielle Validität (Sprachgruppenvergleich; Studie 2) sowie die Kriteriumsvalidität (Intelligenz und weitere Merkmale als Prädiktoren für Schulleistungen; Studie 3 und 4) analysiert; letztere außerdem sowohl prädiktiv

(Intelligenz; Studie 3 und 4) als auch konkurrent (Gewissenhaftigkeit und Leistungsmotivation; Studie 4). Drittens ist positiv hervorzuheben, dass die prädiktive Validität der IDS-1 sowohl über drei (Studie 3) als auch über sieben Jahre (Studie 4) untersucht wurde. Viertens wurde zusätzlich die prädiktive Validität der IDS-1 anderen Intelligenztests gegenübergestellt (Studie 3), und es wurden zusätzliche Prädiktoren für Schulleistungen analysiert (Studie 4). Fünftens wurden sowohl die gesamten Normierungsstichproben (Studie 1) als auch Substichproben der Normierungs- (Studie 4) respektive Normierungs- und Validierungsstichproben (Studie 2) der IDS miteinbezogen. Diese Stichproben sind groß und repräsentativ hinsichtlich zentraler Variablen wie beispielsweise Alter und Geschlecht (Studie 1) respektive vergleichsweise groß (Studie 2 und 4) und weisen breite Altersspannen auf (Studie 1 und 2). Weiter wurden in Studie 2 drei (Monolinguale, Bilinguale und Fremdsprachige) statt wie in anderen Studien nur zwei (Deutsch- und Fremdsprachige) Sprachgruppen einbezogen und diese wurden nach Alter, Geschlecht und mütterlichem Ausbildungsstand gepaart. Studie 4 weist wenig geteilte Methodenvarianz auf, da sowohl ein Test (Intelligenz) als auch Selbstbericht- (Leistungsmotivation) und Fremdbberichtmasse (Gewissenhaftigkeit; Elterneinschätzung) erhoben wurden; zusätzlich wurde die Leistungsmotivation auf Facettenebene erfasst, was im Vergleich zur Faktorebene differenzierte Aussagen ermöglicht. Schließlich wurden in Studie 3 und 4 Schulleistungen in Form von Noten erhoben, welche im Vergleich zu Leistungstests reliabler und von größerer Relevanz sind (Roth et al., 2015).

Nebst den erwähnten Stärken ist eine Limitation dieser Dissertation, dass die Normierungsstichproben der IDS-P und IDS-1 keine Intelligenzminderung und Hochbegabung miteingeschlossen haben und somit keine Generalisierung auf diese Extremgruppen möglich ist (Studie 1). Ebenfalls beziehen sich Studie 3 und 4 auf sich normativ entwickelnde Jugendliche, sodass die Befunde nicht auf andere Altersgruppen und andere Stichproben wie beispielsweise klinische Gruppen generalisierbar sind. Eine weitere Schwäche ist, dass die Sprachgruppen aufgrund fehlender Selbstangaben zu den Deutschkenntnissen eine gewisse Heterogenität aufweisen und dass keine vorgängige Überprüfung der Messinvarianz stattgefunden hat (Studie 2). Ebenfalls ist die hohe Attritionsrate und somit kleine Stichprobengröße zu Messzeitpunkt 2 bei Studie 3 kritisch zu betrachten. Eine weitere Einschränkung ist, dass Gewissenhaftigkeit als Prädiktor für Schulleistungen nur auf Faktorebene untersucht wurde, sodass keine differenzierten Aussagen zu spezifischen Facetten wie im Falle der Leistungsmotivation möglich sind (Studie 4). Für die Facetten der Leistungsmotivation ist wiederum kritisch anzumerken, dass deren Reliabilität teilweise nur knapp ausreichend war (Studie 4)—die Koeffizienten lagen allerdings im Bereich derjenigen der Normierungsstichprobe des FLM 7-13 (vgl. Petermann & Winkel,

2007). Schließlich ist zu erwähnen, dass Studie 3 und 4 im Hinblick auf die Intelligenz zwar ein längsschnittliches Design aufweisen, dass die Resultate jedoch aufgrund fehlender experimenteller Manipulation nicht kausal interpretiert werden dürfen.

6.7 Ausblick

Die vorliegende Dissertation liefert verschiedene Ansatzpunkte für weitere Forschung. Zukünftige Studien sollten Geschlechts- sowie Sprachgruppendifferenzen in verschiedenen Altersgruppen untersuchen. Bezüglich des Geschlechts bestehen Hinweise darauf, dass Unterschiede insbesondere im Jugendalter verstärkt auftreten (Alsaker & Bütikofer, 2005); zudem zeigt die Intelligenz eine Ausdifferenzierung mit dem Alter (Rost, 2009). Hinsichtlich der Sprache findet die Literatur, dass sich sprachliche Defizite langfristig in nichtsprachlichen Bereichen manifestieren und mit zunehmendem Alter größere Diskrepanzen zwischen Deutsch- und Fremdsprachigen bestehen (Daseking et al., 2008).

Weiter sollten Studien die Messinvarianz sowie Geschlechtsdifferenzen in anderen Funktionsbereichen der IDS-P sowie Sprachgruppendifferenzen in anderen Funktionsbereichen der IDS-2 analysieren. Die Literatur zeigt beispielsweise, dass Jungen einen Leistungsvorsprung in der Motorik (Ahnert & Schneider, 2007), Mädchen hingegen einen Vorsprung in sozialen Kompetenzen (Kunter & Stanat, 2002) aufweisen, und dass Fremd- im Vergleich zu Deutschsprachigen in sprachlichen und teilweise mathematischen Fähigkeiten tiefere Werte erreichen (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013).

Schließlich sollten in Bezug auf die Vorhersage von Schulleistungen weitere Aspekte (z.B. Roberts & Wood, 2006) sowie weitere Systeme und Kontexte (z.B. Bronfenbrenner, 1992) berücksichtigt werden: Zum einen sollten Intelligenzfaktoren (z.B. Calvin, Fernandes, Smith, Visscher & Deary, 2010), Gewissenhaftigkeitsfacetten (z.B. Nofle & Robins, 2007), andere Leistungsmotivationsfacetten (z.B. Robbins et al., 2004) und andere Motivationsaspekte (z.B. Steinmayr & Spinath, 2009) miteinbezogen werden. Zum anderen könnten weitere Einflussfaktoren wie beispielsweise elterliche Unterstützung (z.B. Estell & Perdue, 2013), Erziehungs- (z.B. Areepattamannil, 2010) und Unterrichtsstil (z.B. Dever & Karabenick, 2011), Schulverhalten (z.B. Spengler, Damian & Roberts, 2018) oder Beziehungen zu Gleichaltrigen (z.B. Fortuin, van Geel & Vedder, 2016) berücksichtigt werden. Ebenfalls könnten mögliche Mediations- und Moderationseffekte überprüft werden. So gibt es beispielsweise Befunde, dass Gewissenhaftigkeit den Zusammenhang zwischen Intelligenz und Noten (Arbabi, Vollmer, Dörfler & Randler, 2015) respektive Leistungsmotivation den Zusammenhang zwischen Gewissenhaftigkeit und Noten (Richardson & Abraham, 2009) mediiert. Zudem hat sich gezeigt, dass Intelligenz den Zusammenhang zwischen Persönlichkeit und Noten (Heaven & Ciarrochi,

2012) sowie der Testtyp (verbal oder nonverbal) und die Schulstufe den Zusammenhang zwischen Intelligenz und Noten (Roth et al., 2015) moderieren. Weiter sollte zusätzlich zur konkurrenten die prädiktive Validität von Gewissenhaftigkeit und Leistungsmotivation sowie die prädiktive Validität der Intelligenz über einen Zeitraum von mehr als sieben Jahren untersucht werden, um deren lang- respektive längerfristige Einflüsse auf Schulleistungen zu ergründen. Schließlich sollten im Hinblick auf die Vorhersage von Schulleistungen weitere fachspezifische Noten sowie standardisierte Leistungstests als abhängige Variablen eingesetzt werden. Da Leistungstests die Leistung zu einem Zeitpunkt abbilden und weniger von anderen Aspekten wie beispielsweise Selbstkontrolle beeinflusst werden, finden sich in der Regel größere Zusammenhänge mit der Intelligenz als bei Noten (Rost, 2009).

Für alle IDS-Versionen könnten schließlich verschiedene weitere Validitätsanalysen durchgeführt werden; beispielsweise sind für die IDS-P und IDS-2 (repräsentative) Befunde zur prädiktiven Validität gänzlich ausstehend. Da die IDS in anderen Sprachen vorliegen, könnten zudem länderübergreifende Studien gemacht werden. Als Beispiel ließe sich für alle drei deutschsprachigen IDS-Versionen die Faktorenstruktur mit derjenigen der polnischen IDS vergleichen. Da die IDS außerdem sowohl die Intelligenz als auch weitere entwicklungsrelevante Funktionsbereiche erfassen, könnten ebenfalls bereichsübergreifende Analysen durchgeführt werden. Beispielsweise könnte für die IDS-2 die Kriteriumsvalidität von Intelligenz, Gewissenhaftigkeit und Leistungsmotivation in einem Testverfahren analysiert werden. Eine solche funktionsbereichsübergreifende Untersuchung würde weiterführend dazu beitragen, „das Kind in seiner Ganzheit [zu] verstehen und [zu] erfassen“ (Grob, 2018, S. 2).

6.8 Konklusion

Zusammenfassend erweitert die vorliegende Dissertation den Forschungsstand zur Validität der Intelligenz der IDS. Nebst den in Kapitel 2.4 erläuterten Validierungsbefunden weisen die IDS-P und IDS-1 Konstruktvalidität und die IDS-2 differentielle Validität hinsichtlich verschiedener Sprachgruppen auf. Weiter zeigen die IDS-1 im Rahmen der Kriteriumsvalidität prädiktive Validität für Schulleistungen über drei und über sieben Jahre, und die prädiktive Validität über drei Jahre ist mit derjenigen von anderen deutschsprachigen Intelligenztests vergleichbar. In Bezug auf Schulleistungen kann Intelligenz zwar einen bedeutenden Varianzanteil erklären, zusätzlich weisen jedoch weitere Merkmale wie Gewissenhaftigkeit und Leistungsmotivation inkrementelle konkurrente Validität über Intelligenz hinaus auf. Insgesamt kann basierend auf den Ergebnissen dieser Dissertation geschlussfolgert werden, dass die IDS-P, IDS-1 und IDS-2 valide Testverfahren darstellen, die eine gültige Intelligenzdiagnostik im Kindes- und Jugendalter ermöglichen.

LITERATURVERZEICHNIS

- Ahnert, J. & Schneider, W. (2007). Entwicklung und Stabilität motorischer Fähigkeiten vom Vorschul- bis ins frühe Erwachsenenalter. Befunde der Münchner Längsschnittstudie LOGIK. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 39, 12-24. doi: 10.1026/0049-8637.39.1.12
- Alsaker, F. D. & Bütikofer, A. (2005). Geschlechtsunterschiede im Auftreten von psychischen und Verhaltensstörungen im Jugendalter. *Kindheit und Entwicklung*, 14, 169-180. doi: 10.1026/0942-5403.14.3.169
- Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Berlin: Springer. doi: 10.1007/3-540-28507-5
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education & Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arbabi, T., Vollmer, C., Dörfler, T. & Randler, C. (2015). The influence of chronotype and intelligence on academic achievement in primary school is mediated by conscientiousness, midpoint of sleep and motivation. *Chronobiology International*, 32, 349-357. doi: 10.3109/07420528.2014.980508
- Arbuckle, J. L. (2013). *IBM SPSS AMOS (Version 22)*. Chicago, IL: SPSS.
- Arden, R. & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, 41, 39-48. doi: 10.1016/j.paid.2005.11.027
- Areepattamannil, S. (2010). Parenting practices, parenting style, and children's school achievement. *Psychological Studies*, 55, 283-289. doi: 10.1007/s12646-010-0043-0
- Asendorpf, J. B. (1998). *Fünf-Faktoren-Fragebogen für Kinder (FFFK)*. Berlin: Humboldt-Universität, Institut für Psychologie.
- Aspara, J., Wittkowski, K. & Luo, X. (2018). Types of intelligence predict likelihood to get married and stay married: Large-scale empirical evidence for evolutionary theory. *Personality and Individual Differences*, 122, 1-6. doi: 10.1016/j.paid.2017.09.028
- Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36, 189-208. doi: 10.1016/S0021-9924(03)00019-4
- Baltes, P. B. (1990). Entwicklungspsychologie der Lebensspanne: Theoretische Leitsätze. *Psychologische Rundschau*, 41, 1-24.
- Barnett, A. & Stuart, N. (in prep.). *Intelligence and Development Scales – 2 (IDS-2)*. [working title].

- Beauducel, A. & Leue, A. (2014). *Psychologische Diagnostik*. Göttingen: Hogrefe.
- Binet, A. & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191-244. doi: 10.3406/psy.1904.3675
- Börnert, M. & Wilbert, J. (2016). Dynamisches Testen als neue Perspektive in der sonderpädagogischen Diagnostik—Theorie, Evidenzen, Impulse für Forschung und Praxis. *Zeitschrift für Heilpädagogik*, 67, 156-167.
- Brähler, E., Holling, H., Leutner, D. & Petermann, F. (Hrsg.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (3. Aufl.). Göttingen: Hogrefe.
- Bronfenbrenner, U. (1992). Ecological systems theory. In R. Vasta (Ed.), *Six theories of child development* (pp. 187-249). London: Jessica Kingsley.
- Bundesamt für Statistik. (2017). *Migration und Integration*. Verfügbar unter: <https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/migration-integration/auslaendische-bevoelkerung.assetdetail.3302662.html> [01.08.2018].
- Calero, M. D, Fernández-Parra, A., López-Rubio, S., Carles, R., Mata, S., del Carmen Vives, M., . . . Márquez, J. (2013). Variables involved in personal, social and school adjustment in a sample of preschool-aged children from different cultural backgrounds. *European Journal of Psychological Education*, 28, 133-155. doi: 10.1007/s10212-012-0107-8
- Calvin, C. M., Deary, I. J., Fenton, C., Roberts, B. A., Der, G., Leckenby, N. & Batty, G. D. (2011). Intelligence in youth and all-cause-mortality: Systematic review with meta-analysis. *International Journal of Epidemiology*, 40, 626-644. doi: 10.1093/ije/dyq190
- Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M. & Deary, I. J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old school-children in England. *Intelligence*, 38, 424-432. doi: 10.1016/j.intell.2010.04.005
- Camarata, S. & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, 34, 231-252. doi: 10.1016/j.intell.2005.12.001
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511571312
- Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, 38, 592.
- Chen, H. & Zhu, J. (2008). Factor invariance between genders of the Wechsler Intelligence Scale for Children – Fourth edition. *Personality and Individual Differences*, 45, 260-266. doi: 10.1016/j.paid.2008.04.008
- Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers*. Hoboken, NJ: John Wiley & Sons.

- Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9, 233-255. doi: 10.1207/S15328007SEM0902_5
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NY: Erlbaum.
- Costa, P. T. & McCrae, R. R. (1992). *NEO-PI-R Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Daseking, M., Lipsius, M., Petermann, F. & Waldmann, H.-C. (2008). Intelligenz und kulturelle Einflüsse. Differenzen im Intelligenzprofil bei Kindern mit Migrationshintergrund: Befunde zum HAWIK-IV. *Kindheit und Entwicklung*, 17, 76-89. doi: 10.1026/0942-5403.17.2.76
- Daseking, M. & Petermann, F. (2015). Nonverbale Intelligenzdiagnostik: Sprachfreie Erhebung kognitiver Fähigkeiten und Prävention von Entwicklungsrisiken. *Gesundheitswesen*, 77, 791-792. doi: 10.1055/s-0035-1564145
- Deary, I. J. (2009). Introduction to the special issue on cognitive epidemiology. *Intelligence*, 37, 517-519. doi: 10.1016/j.intell.2009.05.001
- Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M. & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish Mental Survey 1932. *Intelligence*, 31, 533-542. doi: 10.1016/S0160-2896(03)00053-9
- Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R. & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence*, 28, 49-55. doi: 10.1016/S0160-2896(99)00031-8
- Dehn, M. J. (2008). *Working memory and academic learning: Assessment and intervention*. Hoboken, NJ: Wiley.
- Deutsches Institut für Normierung. (2002). *DIN 33430—Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Dever, B. V. & Karabenick, S. A. (2011). Is authoritative teaching beneficial for all students? A multi-level model of the effects of teaching style on interest and achievement. *School Psychology Quarterly*, 26, 131-144. doi: 10.1037/a0022985
- Di Domenico, S. I. & Fournier, M. A. (2015). Able, ready, and willing: Examining the additive and interactive effects of intelligence, conscientiousness, and autonomous motivation on undergraduate academic performance. *Learning and Individual Differences*, 40, 156-162. doi: 10.1016/j.lindif.2015.03.016

- Di Fabio, A. & Busoni, L. (2007). Fluid intelligence, personality traits and scholastic success: Empirical evidence in a sample of Italian high school students. *Personality and Individual Differences*, 43, 2095-2104. doi: 10.1016/j.paid.2007.06.025
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135-168. doi: 10.1146/annurev-psych-113011-143750
- Döpfner, M. & Petermann, F. (2008). Diagnostik psychischer Störungen im Kindes- und Jugendalter (3. Aufl.). Göttingen: Hogrefe.
- Dumfart, B. & Neubauer, A. C. (2016). Conscientiousness is the most powerful noncognitive predictor of school achievement in adolescents. *Journal of Individual Differences*, 37, 8-15. doi: 10.1027/1614-0001/a000182
- Eccles, J. S. & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109-132. doi: 10.1146/annurev.psych.53.100901.135153
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1-26. doi: 10.1214/aos/1176344552
- Estell, D. B. & Perdue, N. H. (2013). Social support and behavioral and affective school engagement: The effects of peers, parents, and teachers. *Psychology in the Schools*, 50, 325-339. doi: 10.1002/pits.21681
- Eurostat. (2017). *Migration and migrant population statistics*. Verfügbar unter: http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics [01.08.2018].
- Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, 1, 137-153. doi: 10.1207/S15327574IJT0102_3
- Fecenec, D., Jaworowska, A. & Matczak, A. (2015). *IDS—Skale Inteligencji i Rozwoju dla Dzieci w Wiekach Przedszkolnym*. Warschau: Hogrefe.
- Fecenec, D., Jaworowska, A. & Matczak, A. (in prep.). *Intelligence and Development Scales – 2 (IDS-2)*. [working title].
- Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62, 61-84. doi: 10.3102/00346543062001061
- Fergusson, D. M., Horwood, L. J. & Ridder, E. M. (2005). Show me the child at seven II: Childhood intelligence and later outcomes in adolescence and young adulthood. *Journal of Child Psychology and Psychiatry*, 46, 850-858. doi: 10.1111/j.1469-7610.2005.01472.x

- Ferri, R., Casagrande, M. & Rea, M. (in prep.). *Intelligence and Development Scales – 2 (IDS-2)*. [working title].
- Ferri, R., Rea, M. & Casagrande, M. (2015). *IDS—Intelligence and Development Scales*. Florenz: Hogrefe.
- Flanagan, D. P. & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York, NY: Wiley.
- Fortuin, J., van Geel, M. & Vedder, P. (2016). Peers and academic achievement: A longitudinal study on selection and socialization effects of in-class friends. *Journal of Educational Research, 109*, 1-6. doi: 10.1080/00220671.2014.917257
- Friedman, H. S. & Schustack, M. W. (2004). *Persönlichkeitspsychologie und differentielle Psychologie* (2. Aufl.). München: Pearson.
- Goldbeck, L., Daseking, M., Hellwig-Brida, S., Waldmann, H. C. & Petermann, F. (2010). Sex differences on the German Wechsler Intelligence Test for Children (WISC-IV). *Journal of Individual Differences, 31*, 22-28. doi: 10.1027/1614-0001/a000003
- Goldstein, S., Princiotta, D. & Naglieri, J. A. (Eds.). (2015). *Handbook of intelligence. Evolutionary theory, historical perspective, and current concepts*. New York, NY: Springer. doi: 10.1007/978-1-4939-1562-0
- Gonzalez-Mulé, E., Carter, K. M. & Mount, M. K. (2017). Are smarter people happier? Meta-analyses of the relationships between general mental ability and job and life satisfaction. *Journal of Vocational Behavior, 99*, 146-164. doi: 10.1016/j.jvb.2017.01.003
- Gottfredson, L. S. (1997). Intelligence and social policy [Special issue]. *Intelligence, 24*.
- Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance, 15*, 25-46. doi: 10.1080/08959285.2002.9668082
- Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31*, 343-397. doi: 10.1016/S0160-2896(02)00085-5
- Gottfredson, L. S. & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*, 1-4. doi: 10.1111/j.0963-7214.2004.01301001.x
- Grigorenko, E. L. & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75-111. doi: 10.1037/0033-2909.124.1.75
- Grob, A. (2018). *IDS-2—Intelligenz- und Entwicklungsdiagnostik für die Praxis für Kinder und Jugendliche*. [Präsentation der Informationsveranstaltung in Basel, 14.05.2018].
- Grob, A., Gygi, J. T. & Hagmann-von Arx, P. (in Druck). *The Stanford-Binet Intelligence Scales – Fifth edition (SB5) – Deutsche Version*. Bern: Hogrefe.

- Grob, A. & Hagmann-von Arx, P. (Hrsg.). (2012). *Fallbuch IDS—Die Intelligence and Development Scales in der Praxis*. Göttingen: Hogrefe.
- Grob, A. & Hagmann-von Arx, P. (2018). *Intelligence and Development Scales – 2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche*. Bern: Hogrefe. [Test-manuale: unkorrigierte Vorabexemplare].
- Grob, A., Hagmann-von Arx, P. & Bodmer, N. M. (2009). Entwicklungsdiagnostik. In S. Schneider & J. Margraf (Hrsg.), *Lehrbuch der Verhaltenstherapie, Band 3: Störungen des Kindes- und Jugendalters* (S. 145-158.). Berlin: Springer.
- Grob, A., Meyer, C. S. & Hagmann-von Arx, P. (2009). *Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5-10 Jahren*. Bern: Hans Huber.
- Grob, A., Meyer, C. S. & Hagmann-von Arx, P. (2013). *Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5-10 Jahren* (2. überarb. Aufl.). Bern: Hans Huber.
- Grob, A., Reimann, G., Gut, J. & Frischknecht, M.-C. (2013). *Intelligence and Development Scales – Preschool (IDS-P). Intelligenz- und Entwicklungsskalen für das Vorschulalter*. Bern: Hans Huber.
- Gunderson, L. & Siegel, L. S. (2001). The evils of the use of IQ tests to define learning disabilities in first- and second-language learners. *The Reading Teacher*, 55, 48-55.
- Gut, J., Reimann, G. & Grob, A. (2012). Kognitive, sprachliche, mathematische und sozial-emotionale Kompetenzen als Prädiktoren späterer schulischer Leistungen: Können die Leistungen eines Kindes in den IDS dessen Schulnoten drei Jahre später vorhersagen? *Zeitschrift für Pädagogische Psychologie*, 26, 213-220. doi: 10.1024/1010-0652/a000070
- Gut, J., Reimann, G. & Grob, A. (2013). A contextualized view on long-term predictors of academic performance. *Journal of Educational Psychology*, 105, 436-443. doi: 10.1037/a0031503
- Guthke, J., Beckmann, J. F. & Wiedl, K. H. (2003). Dynamik im dynamischen Testen. *Psychologische Rundschau*, 54, 225-232. doi: 10.1026//0033-3042.54.4.225
- Guthke, J. & Wiedl, K. H. (1996). *Dynamisches Testen*. Göttingen: Hogrefe.
- Häcker, H., Leutner, D. & Amelang, M. (1998). Standards für pädagogisches und psychologisches Testen. *Zeitschrift für Differentielle und Diagnostische Psychologie, Supplementum 1*.

- Hagmann-von Arx, P. & Grob, A. (2014). *Reynolds Intellectual Assessment Scales. Deutschsprachige Adaptation der Reynolds Intellectual Assessment Scales (RIASTM) von Cecil R. Reynolds und Randy W. Kamphaus. Manual und Testset*. Bern: Hans Huber.
- Hagmann-von Arx, P., Grob, A., Petermann, F. & Daseking, M. (2012). Konkurrente Validität des HAWIK-IV und der Intelligence and Development Scales (IDS). *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 40, 41-50. doi: 10.1024/1422-4917/a000148
- Hagmann-von Arx, P., Lemola, S. & Grob, A. (2016). Does IQ = IQ? Comparability of intelligence test scores in typically developing children. *Assessment*, 1-11. doi: 10.1177/1073191116662911
- Hagmann-von Arx, P., Meyer, C. S. & Grob, A. (2008a). Assessing intellectual giftedness with the WISC-IV and the IDS. *Zeitschrift für Psychologie / Journal of Psychology*, 216, 172-179. doi: 10.1027/0044-3409.216.3.172
- Hagmann-von Arx, P., Meyer, C. S. & Grob, A. (2008b). Intelligenz- und Entwicklungsdiagnostik im deutschen Sprachraum. *Kindheit und Entwicklung*, 17, 232-242. doi: 10.1026/0942-5403.17.4.232
- Hagmann-von Arx, P., Petermann, F. & Grob, A. (2013). Konvergente und diskriminante Validität der WISC-IV und der Intelligence and Development Scales (IDS) bei Kindern mit Migrationshintergrund. *Diagnostica*, 59, 170-182. doi: 10.1026/0012-1924/a000091
- Heaven, P. C. L. & Ciarrochi, J. (2012). When IQ is not everything: Intelligence, personality and academic performance at school. *Personality and Individual Differences*, 53, 518-522. doi: 10.1016/j.paid.2012.04.024
- Heaven, P. C. L., Ciarrochi, J. & Vialle, W. (2007). Conscientiousness and Eysenckian psychoticism as predictors of school grades: A one-year longitudinal study. *Personality and Individual Differences*, 42, 535-546. doi: 10.1016/j.paid.2006.07.028
- Hessels, M. G. P. (1997). Low IQ but high learning potential: Why Zeyneb and Moussa do not belong in special education. *Educational and Child Psychology*, 14, 121-136.
- Holling, H., Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383-386.
- Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities* (Unpublished doctoral dissertation). University of Illinois, Champaign, IL.

- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57, 253-270. doi: 10.1037/h0023816
- Howitt, D. & Cramer, D. (2008). *Introduction to research methods in psychology* (2nd ed.). Harlow: Pearson Education Limited.
- Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55. doi: 10.1080/10705519909540118
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581-592. doi: 10.1037/0003-066X.60.6.581
- Hyde, J. S. & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, 106, 8801-8807. doi: 10.1073/pnas.0901265106
- IBM. (2012). *SPSS Statistics for Mac (Version 22)*. Armonk, NY: IBM Corp.
- IBM. (2013). *SPSS Statistics for Mac (Version 23)*. Armonk, NY: IBM Corp.
- Janke, N., Daseking, M. & Petermann, F. (2008). Intelligenzdiagnostik im Kindergartenalter—Ein Beitrag zur Validierung des SON-R 2½-7. *Diagnostica*, 54, 174-183. doi: 10.1026/0012-1924.54.4.174
- Jaworowska, A., Matczak, A. & Fecenec, D. (2011). *IDS—Skale Inteligencji i Rozwoju dla Dzieci w wieku 5-10 lat*. Warschau: Hogrefe.
- Johnson, W. & Bouchard, T. J. J. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence*, 35, 23-39. doi: 10.1016/j.intell.2006.03.012
- Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl.) (S. 27-74). Berlin: Springer. doi: 10.1007/978-3-642-20072-4
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342. doi: 10.1111/j.1745-3984.2001.tb01130.x
- Kaufman, A. S., Raiford, S. E. & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Hoboken, NJ: Wiley.
- Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R. & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children – Fourth edition: What does it measure? *School Psychology Review*, 35, 108-127.

- Keith, T. Z., Reynolds, M. R., Patel, P. G. & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock-Johnson III Tests of Cognitive Abilities. *Intelligence*, 36, 502-525. doi: 10.1016/j.intell.2007.11.001
- Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L. & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales – Second edition. *Intelligence*, 39, 389-404. doi: 10.1016/j.intell.2011.06.008
- Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau*, 57, 243-253. doi: 10.1026/0033-3042.57.4.243
- Kramer, J. (1972). *Intelligenztest: Mit einer Einführung in Theorie und Praxis der Intelligenzprüfung* (4. Aufl.). Solothurn: Antonius.
- Krapp, A. (1997). Selbstkonzept und Leistung—Dynamik ihres Zusammenspiels: Literaturüberblick. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 325-339). Weinheim: Beltz.
- Krejčířová, D., Urbánek, T., Širůček, J. & Jabůrek, M. (2013). *IDS—Intelligenční a vývojová škála pro děti ve věku 5-10 let Prague*. Prag: Hogrefe.
- Kunter, M. & Stanat, P. (2002). Soziale Kompetenz von Schülerinnen und Schülern. *Zeitschrift für Erziehungswissenschaft*, 5, 49-71. doi: 10.1007/s11618-002-0004-z
- Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, 17, 257-271. doi: 10.1016/0191-8869(94)90030-2
- Maccoby, E. E. & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-181). New York, NY: Guilford Press.
- McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.) (pp. 136-182). New York, NY: Guilford Press.
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1-10. doi: 10.1016/j.intell.2008.08.004

- Meyer, C. S., Hagmann-von Arx, P. & Grob, A. (2009). Die Intelligence and Development Scale Sozial-Emotionale Kompetenz (IDS-SEK): Psychometrische Eigenschaften eines Tests zur Erfassung sozial-emotionaler Fähigkeiten. *Diagnostica*, 55, 234-244. doi: 10.1026/0012-1924.55.4.234
- Meyer, C. S., Hagmann-von Arx, P., Lemola, S. & Grob, A. (2010). Correspondence between the general ability to discriminate sensory stimuli and general intelligence. *Journal of Individual Differences*, 31, 46-56. doi: 10.1027/1614-0001/a000006
- Milfont, T. L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural-research. *International Journal of Psychological Research*, 3, 111-121. doi: 10.21500/20112084.857
- Millsap, R. E. & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93-115. doi: 10.1037/1082-989X.9.1.93
- Moosbrugger, H. & Höfling, V. (2006). Teststandards. In F. Petermann & M. Eid (Hrsg.), *Handbuch der psychologischen Diagnostik* (S. 407-419). Göttingen: Hogrefe.
- Moosbrugger, H. & Höfling, V. (2012). Standards für psychologisches Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl.) (S. 203-224). Berlin: Springer. doi: 10.1007/978-3-642-20072-4
- Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl.) (S. 7-26). Berlin: Springer. doi: 10.1007/978-3-642-20072-4
- Noftle, E. E. & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93, 116-130. doi: 10.1037/0022-3514.93.1.116
- O'Connor, M. C. & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, 43, 971-990. doi: 10.1016/j.paid.2007.03.017
- Palejwala, M. H. & Fine, J. G. (2015). Gender differences in latent cognitive abilities in children aged 2 to 7. *Intelligence*, 48, 96-108. doi: 10.1016/j.intell.2014.11.004
- Pawlik, K. (1968). *Dimensionen des Verhaltens*. Bern: Hans Huber.
- Petermann, F. (2017). *Wechsler Intelligence Scale for Children – Fifth edition (WISC-V) – Deutsche Version*. Frankfurt a. M.: Pearson.
- Petermann, F. & Petermann, U. (2011). *Wechsler Intelligence Scale for Children – Fourth edition (WISC-IV) – Deutsche Version*. Frankfurt a. M.: Pearson.

- Petermann, F. & Winkel, S. (2007). *Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse (FLM 7-13)*. Frankfurt a. M.: Harcourt Test Services.
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135, 322-338. doi: 10.1037/a0014996
- Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. *British Journal of Educational Psychology*, 84, 239-252. doi: 10.1111/bjep.12019
- Qualter, P., Gardner, K. J., Pope, D. J., Hutchinson, J. M. & Whiteley, H. E. (2012). Ability emotional intelligence, trait emotional intelligence, and academic success in British secondary schools: A 5 year longitudinal study. *Learning and Individual Differences*, 22, 83-91. doi: 10.1016/j.lindif.2011.11.007
- R Core Team. (2016). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.
- Reich, H., Roth, H.-J., Dirim, I., Jörgensen, J. N., List, G., Neumann, U., . . . Wurnig, V. (Hrsg.). (2002). *Spracherwerb zweisprachig aufwachsender Kinder und Jugendlicher: Ein Überblick über den Stand der nationalen und internationalen Forschung*. Hamburg: Behörde für Bildung und Sport.
- Reynolds, C. R. & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales (RIAS): Professional Manual*. Lutz, FL: PAR.
- Reynolds, M. R., Keith, T. Z., Ridley, K. P. & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, 36, 236-260. doi: 10.1016/j.intell.2007.06.003
- Richardson, M. & Abraham, C. (2009). Conscientiousness and achievement motivation predict performance. *European Journal of Personality*, 23, 589-605. doi: 10.1002/per.732
- Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138, 353-387. doi: 10.1037/a0026838
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261-288. doi: 10.1037/0033-2909.130.2.261
- Roberts, B. W. & Wood, D. (2006). Personality development in the context of the neo-socio-analytic model of personality. In D. Mroczek & T. Little (Eds.), *Handbook of personality development* (pp. 11-39). Mahwah, NJ: Erlbaum. doi: 10.4324/9781315805610.ch2

- Rosander, P. & Bäckström, M. (2014). Personality traits measured at baseline can predict academic performance in upper secondary school three years later. *Scandinavian Journal of Psychology*, 55, 611-618. doi: 10.1111/sjop.12165
- Rost, D. H. (2009). *Intelligenz: Fakten und Mythen*. Weinheim: Beltz.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F. & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137. doi: 10.1016/j.intell.2015.09.002
- Ruch, F. L. & Zimbardo, P. G. (2013). *Lehrbuch der Psychologie: Eine Einführung für Studenten der Psychologie, Medizin und Pädagogik*. Berlin: Springer.
- Ruiter, S., Geerds, F., Timmerman, M. & Visser, L. (in prep.). *Intelligence and Development Scales – 2 (IDS-2)*. [working title].
- Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences*, 53, 137-141. doi: 10.1016/j.paid.2011.06.013
- Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, 49, 304-313. doi: 10.1037/0003-066X.49.4.304
- Schmidt, F. L. & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86, 162-173. doi: 10.1037/0022-3514.86.1.162
- Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin: Springer. doi: 10.1007/978-3-642-17001-0
- Schneider, W. J. & Flanagan, D. P. (2015). The relationship between theories of intelligence and intelligence tests. In S. Goldstein, D. Princiotta & J. A. Naglieri (Eds.), *Handbook of intelligence. Evolutionary theory, historical perspective, and current concepts* (pp. 317-340). New York, NY: Springer. doi: 10.1007/978-1-4939-1562-0_21
- Sellers, A. H., Burns, W. J. & Guyrke, J. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence – Revised as a function of stratification variables. *Applied Neuropsychology*, 9, 65-73. doi: 10.1207/S15324826AN0902_1
- Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology*, 15, 201-292. doi: 10.2307/1412107
- Spengler, M., Damian, R. I. & Roberts, B. W. (2018). How you behave in school predicts life success above and beyond family background, broad traits, and cognitive ability. *Journal of Personality and Social Psychology*, 114, 620-636. doi: 10.1037/pspp0000185

- Spinath, B., Freudenthaler, H. & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences*, 48, 481-486. doi: 10.1016/j.paid.2009.11.028
- Steinmayr, R. & Spinath, B. (2007). Predicting school achievement from motivation and personality. *Zeitschrift für Pädagogische Psychologie*, 21, 207-216. doi: 10.1024/1010-0652.21.3.207
- Steinmayr, R. & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, 19, 80-90. doi: 10.1016/j.lindif.2008.05.004
- Stern, E. & Neubauer, A. (2016). Intelligenz: Kein Mythos, sondern Realität. *Psychologische Rundschau*, 67, 1-13. doi: 10.1026/0033-3042/a000290
- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, 35, 401-426. doi: 10.1016/j.intell.2006.09.004
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2007). *Snijders Oomen nonverbaler Intelligenztest 2½-7 (SON-R 2½-7)*. Göttingen: Hogrefe.
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2012). *Snijders Oomen nonverbaler Intelligenztest 6-40 (SON-R 6-40)*. Göttingen: Hogrefe.
- Testkuratorium. (2010). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologinnenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau*, 61, 52-56.
- Toussaint, A., Heinze, L., Lipsius, M. & Petermann, F. (2012). Zur Aussagekraft des SON-R 6-40 bei Kindern mit Hörbeeinträchtigung und Kindern mit Migrationshintergrund. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 61, 108-121. doi: 10.13109/prkk.2012.61.2.108
- Van de Schoot, R, Lugtig, P. & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9, 486-492. doi: 10.1080/17405629.2012.686740
- Weiss, L. G., Harris, J. G., Prifitera, A., Courville, T., Rolfhus, E., Saklofske, D. H. & Holdnack, J. A (2006). WISC-IV interpretation in societal context. The essentials and beyond. In L. G. Weiss, D. H. Saklofske, A. Prifitera & J. A. Holdnack (Eds.), *WISC-IV advanced clinical interpretation* (pp. 1-49). London: Elsevier.
- Weiss, R. H. (2006). *Grundintelligenztest Skala 2 mit Wortschatztest und Zahlenfolgentest – Revision (CFT 20-R mit WS/ZF-R)*. Göttingen: Hogrefe.

- Wigfield, A. & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68-81. doi: 10.1006/ceps.1999.1015
- Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U. & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, 33, 292-296. doi: 10.1037/a0030727

APPENDIX A: Studie 1

Schweizer, F., Hagmann-von Arx, P., Ledermann, T. & Grob, A. (2018). Geschlechtsinvarianz und Geschlechtsdifferenzen in der Intelligenzeinschätzung mit den Intelligence and Development Scales. *Diagnostica*. Eingereichte Manuskriptform; in dieser Form noch nicht für eine Veröffentlichung angenommen [23.10.2017].

Online-Vorveröffentlichung als Schweizer, F., Hagmann-von Arx, P., Ledermann, T. & Grob, A. (2018). Geschlechtsinvarianz und Geschlechtsdifferenzen in der Intelligenzeinschätzung mit den Intelligence and Development Scales. *Diagnostica*. Online-Vorveröffentlichung. doi: 10.1026/0012-1924/a000207 einsehbar.

Geschlechtsinvarianz und -differenzen in den IDS

Titel:

Geschlechtsinvarianz und Geschlechtsdifferenzen in der Intelligenzeinschätzung mit den Intelligence and Development Scales

Kurztitel:

Geschlechtsinvarianz und -differenzen in den IDS

Englischer Titel:

Gender invariance and gender differences in intelligence assessment with the Intelligence and Development Scales

Autoren:

Florine Schweizer¹, Priska Hagmann-von Arx¹, Thomas Ledermann² und Alexander Grob¹

¹ Universität Basel, Fakultät für Psychologie, Missionsstrasse 62, 4055 Basel, Schweiz

² Utah State University, Department of Psychology, Old Main Hill, Logan, Utah 84322, USA

Geschlechtsinvarianz und -differenzen in den IDS

Zusammenfassung

Die Studie untersucht die Messinvarianz über das Geschlecht und Geschlechtsdifferenzen in der Intelligenz mit den *Intelligence and Development Scales–Preschool* (IDS-P) und *Intelligence and Development Scales* (IDS). Die IDS-P (3;0–5;11) und IDS (5;0–10;11) erfassen die allgemeine Intelligenz anhand von sieben Untertests. Die Studie verwendete die Normierungsstichproben mit $N(\text{IDS-P})=700$ und $N(\text{IDS})=1330$ (je 50% Jungen). Beide Verfahren wiesen latente partielle Messinvarianz auf Ebene der Untertests sowie auf Ebene der allgemeinen Intelligenz auf. Auf manifester Ebene zeigten sich keine Mittelwertdifferenzen für die allgemeine Intelligenz. In einigen Untertests ergaben sich Unterschiede zwischen den Geschlechtern: Jungen zeigen im Durchschnitt in *Denken Bildlich* in beiden Verfahren, Mädchen in *Aufmerksamkeit Selektiv* und *Gedächtnis Auditiv* in den IDS-P bessere Leistungen. Die manifesten Varianzen unterschieden sich nicht. Die Befunde belegen, dass die Verfahren invariant bezüglich ihrer Struktur über das Geschlecht sind und folglich eine faire Intelligenzeinschätzung für beide Geschlechter ermöglichen.

Schlüsselwörter: Geschlechtsinvarianz; Geschlechtsdifferenzen; Intelligenz; IDS-P; IDS; Intelligence and Development Scales

Geschlechtsinvarianz und -differenzen in den IDS

Abstract

This study investigates measurement invariance across gender and gender differences in intelligence with the *Intelligence and Development Scales–Preschool* (IDS-P) and *Intelligence and Development Scales* (IDS). The IDS-P (3;0–5;11) and IDS (5;0–10;11) assess general intelligence with seven subtests. This study used the standardization samples with $N(\text{IDS-P})=700$ and $N(\text{IDS})=1330$ (50% boys each). Both scales revealed latent partial measurement invariance on a subtest level and on a general intelligence level. On a manifest level, there were no mean level differences. Some differences occurred in the subtests: Boys scored higher on *Abstract Reasoning* in both scales, girls scored higher on *Selective Attention* and *Long-Term Memory* in the IDS-P. Additionally, no manifest variance differences were found. Results indicate that both scales hold measurement invariance across gender, and therefore provide a fair assessment of intelligence for both boys and girls.

Key words: gender invariance; gender differences; intelligence; IDS-P; IDS; Intelligence and Development Scales

Geschlechtsinvarianz und -differenzen in den IDS

Einleitung

In der Kindheit erfasste Intelligenztestwerte sind prädiktiv für den weiteren Lebenserfolg wie beispielsweise Schulleistung, sozio-ökonomischer Status, physische und psychische Gesundheit sowie Langlebigkeit (Cramer, 2012; Roth et al., 2015; Strenze, 2007; Wrulich et al., 2014). Die Intelligenzeinschätzung stellt somit einen wichtigen Bestandteil in diagnostischen Situationen dar und wird beispielsweise bei der Befunderhebung von Entwicklungs- sowie Verhaltensstörungen oder für die Schultypwahl herangezogen. Folglich gehören Intelligenztests zu den am häufigsten eingesetzten Testverfahren in der psychologischen Praxis (Goldstein, Princiotta & Naglieri, 2015).

Intelligenztests liegen verschiedene Theorien zu Grunde. Gemäß Spearman (1904) Generalfaktor-Theorie ist die Intelligenz hierarchisch aufgebaut und besteht aus der allgemeinen Intelligenz *g* an der Spitze und mehreren untergeordneten spezifischen Faktoren *s*. Cattells (1941) Gf-Gc-Theorie geht von zwei Faktoren aus: fluide Intelligenz (Fähigkeit zur Problemlösung, ohne dass erworbenes Wissen eine bedeutsame Rolle spielt; *Gf*) und kristalline Intelligenz (Fähigkeit, erworbenes Wissen zur Lösung von Problemen anzuwenden; *Gc*). Horn (1965) erweiterte diese Theorie um acht Faktoren. Die Drei-Schichten-Theorie von Carroll (1993) setzt sich aus drei Schichten zusammen: generelle, breite und enge Fähigkeiten. Diese Theorien wurden von McGrew (2005) zur Cattell-Horn-Carroll-Theorie (CHC-Theorie) integriert, welche aktuell die umfassendste und empirisch am besten gestützte Intelligenztheorie darstellt. Die CHC-Theorie besagt, dass die Intelligenz hierarchisch aufgebaut ist und aus drei Schichten besteht. An der Spitze (Stratum III) steht die allgemeine Intelligenz *g*, auf Stratum II befinden sich zehn breite Fähigkeiten. Geordnet nach ihrer Bedeutung für *g* sind dies *Fluide Intelligenz, Mengen- und Zahlenwissen, Kristalline Intelligenz, Lesen und Schreiben, Kurzzeitgedächtnis, Visuelle Wahrnehmung, Auditive Wahrnehmung, Langzeitgedächtnis, Verarbeitungsgeschwindigkeit* und *Reaktions-/Entscheidungszeit*.

Geschlechtsinvarianz und -differenzen in den IDS

Auf der untersten Ebene (Stratum I) finden sich schließlich über 70 enge Fähigkeiten, die mit spezifischen Leistungen in Untertests erhoben werden, wie beispielsweise Wortschatz und Grammatik als Teilaspekte der kristallinen Intelligenz (McGrew, 2005). Aktuelle Intelligenztestverfahren, die auf der CHC-Theorie beruhen, sind beispielsweise die *Wechsler-Skalen* (z.B. Wechsler, 2003).

Die Intelligenz gilt als relativ stabile Persönlichkeitseigenschaft, gleichwohl finden Veränderungen über die Lebensspanne statt (Rost, 2009). In den ersten Lebensjahren kommt es zu einer Differenzierung der Intelligenzstruktur (Abnahme von *g* und Ausbildung von spezifischen Faktoren), während im höheren Erwachsenenalter eine Dedifferenzierung zu beobachten ist (Zunahme von *g*). Ab dem Schuleintritt besteht eine relativ hohe und ab einem Alter von etwa zehn Jahren eine hohe Stabilität, sodass ab Schuleintritt eine Prognose für die Grundschulzeit und ab etwa zehn Jahren auch längerfristige Prognosen für die Entwicklung gestellt werden können (Rost, 2009). Hinsichtlich Niveaustabilität zeigt die Intelligenz eine stetige Zunahme ab Geburt, mit zunehmendem Alter flacht der Anstieg jedoch ab und um das Alter von 25 Jahren herum wird ein Plateau erreicht, während im höheren Alter schließlich eine physiologisch bedingte Abnahme stattfindet (Schaie, 1994). Im Hinblick auf Rangordnungsstabilität weist die Intelligenz ab etwa zehn Jahren bis ins hohe Erwachsenenalter eine hohe Stabilität auf mit Korrelationen um .70 (Deary, Whalley, Lemmon, Crawford & Starr, 2000).

Verschiedene Theorien postulieren Ähnlichkeiten und Differenzen zwischen den Geschlechtern in verschiedenen Intelligenzaspekten. Während die *Geschlechtsähnlichkeitshypothese* (Hyde, 2005) besagt, dass sich die Geschlechter in ihrer Intelligenz ähnlich sind, postuliert die *Differenzhypothese* (Maccoby & Jacklin, 1974), dass Unterschiede zwischen den Geschlechtern bestehen, nämlich Stärken zu Gunsten von Männern in visuell-räumlichen und mathematischen Fähigkeiten und Stärken zu Gunsten von Frauen in verbalen

Geschlechtsinvarianz und -differenzen in den IDS

Fähigkeiten. Gemäß der *Hypothese der größeren männlichen Variabilität* (Hyde & Mertz, 2009) weisen Männer eine größere Variabilität auf, sie zeigen eine breitere Streuung und sind an den Extremen der Normalverteilung (Intelligenzminderung und Hochbegabung) stärker vertreten. Lynn (1994) beschreibt in seiner Entwicklungstheorie Unterschiede in Abhängigkeit des Alters. Die Theorie besagt, dass eine Korrelation zwischen Gehirngröße und Intelligenz besteht, die Geschlechter zu verschiedenen Zeitpunkten eine pubertätsbedingte (Gehirn-)Reifung durchlaufen und sich Unterschiede synchron zur Gehirnentwicklung herausbilden. Jungen und Mädchen weisen bis zu einem Alter von etwa acht Jahren eine vergleichbare Intelligenz auf, ab etwa acht Jahren zeigen Mädchen höhere Mittelwerte aufgrund ihrer früher einsetzenden Pubertät und ab etwa 15 Jahren zeigen schließlich Jungen höhere Mittelwerte, da sie die Mädchen in ihrer Entwicklung einholen. Im Erwachsenenalter liegt schließlich das Maximum dieser Differenzen mit etwa vier Intelligenzpunkten vor. Die *Maskierungshypothese* (Johnson & Bouchard, 2007) schließlich besagt, dass die allgemeine Intelligenz Mittelwertunterschiede in spezifischen Intelligenzfaktoren maskiert. Einerseits kann der Fall bestehen, dass keine Unterschiede in der allgemeinen Intelligenz bestehen, es allerdings Unterschiede in spezifischen Faktoren gibt und diese somit unterschätzt werden, wenn sie zur allgemeinen Intelligenz aufsummiert werden. Andererseits kann der Fall bestehen, dass Unterschiede in der allgemeinen Intelligenz nur durch Unterschiede in spezifischen Faktoren zustande kommen.

Hinsichtlich des Testgütekriteriums Testfairness (Moosbrugger & Kelava, 2008) sind Intelligenztestverfahren so konzipiert, dass sie keine Geschlechtsdifferenzen aufzeigen sollten, um beiden Geschlechtern eine gerechte Einschätzung zu ermöglichen (Rost, 2009). Forschungsbefunde zum Kindes- und Jugendalter zeigen, dass in Normierungsstichproben von Intelligenztestverfahren wie der Deutschen Version der *Wechsler Intelligence Scales for Children–Vierte Auflage* (WISC-IV; Petermann & Petermann, 2011) sowie vergleichbaren

Geschlechtsinvarianz und -differenzen in den IDS

Verfahren in standardisierten Testwerten in der allgemeinen Intelligenz keine Geschlechtsdifferenzen bezüglich Mittelwerten bestehen (Camarata & Woodcock, 2006; Goldbeck, Daseking, Hellwig-Brida, Waldmann & Petermann, 2010; Savage-McGlynn, 2012; Sellers, Burns & Guyrke, 2002). Hingegen gibt es Befunde aus Studien zum Kindes- und Jugendalter, die nicht auf Normierungsdaten basieren, die darauf hinweisen, dass die Varianz in der allgemeinen Intelligenz bei Jungen größer ist (Arden & Plomin, 2006; Deary, Thorpe, Wilson, Starr & Whalley, 2003). In spezifischen breiten und engen Intelligenzaspekten werden im Vergleich zur allgemeinen Intelligenz hingegen geschlechtsspezifische Unterschiede bei Kindern und Jugendlichen deutlich. Verschiedene Normierungsstudien zeigen, dass Jungen in der fluiden Intelligenz und im quantitativen Schlussfolgern (Goldbeck et al., 2010; Keith, Reynolds, Patel & Ridley, 2008) sowie in visuell-räumlichen Fähigkeiten (Reynolds, Keith, Ridley & Patel, 2008) bessere Leistungen erreichen. Mädchen hingegen erzielen bessere Leistungen im Kurz- und Langzeitgedächtnis (Keith et al., 2011) sowie in der Verarbeitungsgeschwindigkeit (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Keith et al., 2008; Keith, Reynolds, Roberts, Winter & Austin, 2011). Im Hinblick auf Varianzen in spezifischen Intelligenzaspekten weisen die Normierungsstudien von Palejwala und Fine (2015) und Reynolds et al. (2008) auf keine signifikanten Geschlechtsunterschiede im Alter zwischen zwei und 18 Jahren hin.

Die Untersuchung von Geschlechtsdifferenzen setzt voraus, dass ein Verfahren sogenannte Messinvarianz aufweist. Milfont und Fischer (2010) unterscheiden vier Arten von Invarianz, welche aufeinander aufbauen: Die konfigurale Invarianz (auch Forminvarianz) ist die schwächste Form der Invarianz und testet die Vergleichbarkeit der Faktorenstruktur über die Gruppen hinweg. Die metrische (auch schwache) Invarianz bedingt, dass die Faktorladungen für die zu vergleichenden Gruppen gleich sind, und impliziert, dass die Konstrukte in allen Gruppen die gleiche inhaltliche Bedeutung haben. Die skalare (auch starke) Invarianz

Geschlechtseinvarianz und -differenzen in den IDS

setzt voraus, dass sich die Gruppen in den y-Achsenabschnitten (Ordinatenabschnitten) der Indikatoren nicht unterscheiden, und geht davon aus, dass in allen Gruppen der gleiche Zusammenhang zwischen den Testitems und dem zu messenden Merkmal besteht. Die Invarianz der Fehlervarianzen (auch strikte Varianz) erfordert, dass es keine Unterschiede zwischen den Gruppen bezüglich den Fehlervarianzen gibt. Bei skalarer Invarianz kann davon ausgegangen werden, dass bei beiden Geschlechtern das gleiche Konstrukt äquivalent gemessen wird (Millsap & Kwok, 2004), was eine Voraussetzung für den Vergleich von latenten Mittelwertunterschieden zwischen Gruppen darstellt (Meredith, 1993). Nur wenn diese Voraussetzung erfüllt ist, kann strukturelle Invarianz untersucht werden. Diese prüft die Invarianz von Varianzen, Kovarianzen und Mittelwerten der latenten Variablen (Milfont & Fischer, 2010). Da vollständige Invarianz in der Praxis kaum identifiziert werden kann, wird als Alternative häufig partielle Invarianz untersucht. Hierbei werden bei der zu überprüfenden Faktorenstruktur nur ausgewählte Parameter gleichgesetzt (Byrne, Shavelson & Muthen, 1989).

Die vorliegende Studie untersucht bei zwei im deutschen Sprachraum häufig eingesetzten Intelligenztestverfahren für Kinder, ob die Verfahren Messinvarianz auf Ebene der Untertests sowie auf Ebene der allgemeinen Intelligenz bezüglich des Alters und des Geschlechts aufweisen und inwiefern sich die Geschlechter in der allgemeinen Intelligenz und in spezifischen Intelligenzaspekten voneinander unterscheiden. Die beiden Testverfahren sind die *Intelligence and Development Scales–Preschool* (IDS-P) für Kinder von 3;0 bis 5;11 Jahren (Grob, Reimann, Gut & Frischknecht, 2013) und die *Intelligence and Development Scales* (IDS) für Kinder von 5;0 bis 10;11 Jahren (Grob, Meyer & Hagmann-von Arx, 2013). Die IDS entsprechen einer Neukonzipierung des Kramer Intelligenztests (KIT; Kramer, 1972), welcher wiederum auf den ersten Intelligenztest von Binet und Simon (1905) zurückgeht. Die IDS-P stellen eine Weiterentwicklung der IDS für Vorschulkinder dar. Die

Geschlechtsinvarianz und -differenzen in den IDS

in den IDS-P und IDS eingeschätzte Intelligenzleistung basiert nicht auf einer spezifischen Intelligenztheorie, ermöglicht jedoch die Einschätzung der allgemeinen Intelligenz nach Spearman (1904) und umfasst insbesondere fluide Intelligenzaspekte (Hagmann-von Arx, Grob, Petermann & Daseking, 2012). Beide Verfahren setzen sich aus den strukturgleichen sieben Untertests zusammen, die sich jeweils verschiedenen Fähigkeiten der CHC-Theorie (McGrew, 2005) zuordnen lassen, diese aber nicht umfassend erfassen: Der Untertest *Wahrnehmung Visuell* erfasst visuelle Wahrnehmung und Diskriminationsfähigkeiten und somit Aspekte der *Visuellen Wahrnehmung* der CHC-Theorie. *Aufmerksamkeit Selektiv* erfordert Verarbeitungsgeschwindigkeit und misst fokussierte Aufmerksamkeit auf bestimmte Merkmale bei gleichzeitiger Inhibition von Störreizen unter Zeitdruck, was der breiten Fähigkeit *Verarbeitungsgeschwindigkeit* zuzuordnen ist. *Gedächtnis Phonologisch* erfasst das Kurzzeitgedächtnis für verbale Information und *Gedächtnis Räumlich-Visuell* für visuell-räumliche Information im Sinne der Fähigkeit *Kurzzeitgedächtnis*. *Gedächtnis Auditiv* misst verbales Langzeitgedächtnis im Sinne der Fähigkeit *Langzeitgedächtnis*. *Denken Bildlich* stellt eine Kombination aus Aspekten der Fähigkeiten *Visuelle Wahrnehmung* und *Fluide Intelligenz* dar und erfasst problemlösendes Denken mit visuellem Gehalt. *Denken Konzeptuell* schließlich erfordert sowohl Aspekte der Fähigkeiten *Fluide Intelligenz* als auch *Kristalline Intelligenz* und erfasst die Einschätzung von bedeutungsbezogenen Wissensrepräsentationen. Zusätzlich zur Intelligenz ermöglichen die IDS-P und IDS eine Einschätzung der allgemeinen Entwicklung in fünf verschiedenen Funktionsbereichen: *Psychomotorik*, *Sozial-Emotionale Kompetenz*, *Denken Logisch-Mathematisch*, *Sprache* und *Umgang mit der Test-situation* in den IDS-P respektive *Leistungsmotivation* in den IDS.

Für die IDS wurden umfangreiche Studien zur Validität durchgeführt. Im Rahmen der Konstruktvalidität wurde die faktorielle Validität überprüft, welche für die kognitiven Untertests als Indikatoren für die allgemeine Intelligenz standardisierte Faktorladungen zwi-

Geschlechtsinvarianz und -differenzen in den IDS

schen .26 und .49 ergab (Meyer, Hagmann-von Arx, Lemola & Grob, 2010). Die konvergente Validität wird durch hohe Zusammenhänge ($r=.54-.83$) zwischen den kognitiven Untertests und den WISC-IV-Indizes unterstützt, während die divergente Validität durch niedrige Zusammenhänge ($r=.03-.34$) zwischen den nichtkognitiven Untertests und den WISC-IV-Indizes bekräftigt wird (Hagmann-von Arx et al., 2012). Des Weiteren wurde die differentielle Validität anhand von verschiedenen Vergleichen nachgewiesen. Es zeigte sich, dass Hochbegabte in den kognitiven Untertests besser abschneiden als durchschnittlich Begabte (Hagmann-von Arx, Meyer & Grob, 2008), dass deutschsprachige Kinder in den sprachbasierten Untertests besser abschneiden als fremdsprachige Kinder (Hagmann-von Arx, Petermann & Grob, 2013) und dass unauffällige Kinder im Funktionsbereich Sozial-Emotionale Kompetenz besser abschneiden als verhaltensauffällige Kinder sowie Kinder mit Asperger-Syndrom (Meyer, Hagmann-von Arx & Grob, 2009). Die IDS weisen außerdem prädiktive Validität auf, indem sie zwischen 9 und 37% Varianz an drei Jahre später erhobenen Schulnoten (Gesamtdurchschnitt, Deutsch, Mathematik und Sachkunde) erklären können ($\Delta R^2=.09-.37$; $\beta=.18-.38$; Gut, Reimann & Grob, 2012; Gygi, Hagmann-von Arx, Schweizer & Grob, 2017). Es liegen allerdings für beide Verfahren bislang keine Befunde zur Messinvarianz und zu möglichen Geschlechtsunterschieden vor.

Forschungsziele und Hypothesen

Das Ziel dieses Artikels ist es, für die IDS-P und IDS in den Normierungsstichproben bei Kindern zwischen drei und zehn Jahren die Messinvarianz über das Alter auf Ebene der Untertests und der allgemeinen Intelligenz zu prüfen. Bei gegebener skalarer Invarianz wird in einem weiteren Schritt die Messinvarianz über das Geschlecht auf Ebene der Untertests und der allgemeinen Intelligenz geprüft. Bei gegebener skalarer Invarianz werden darüber hinaus manifeste Mittelwert- und Varianzdifferenzen in der allgemeinen Intelligenz und in den Untertests untersucht. In Anlehnung an die oben dargestellten Befunde erwarten wir in

Geschlechtsinvarianz und -differenzen in den IDS

Bezug auf Leistungsdifferenzen, dass in der allgemeinen Intelligenz keine Mittelwertunterschiede zwischen den Geschlechtern, jedoch eine größere Varianz bei Jungen beständen.

Bezüglich spezifischen Intelligenzaspekten erwarten wir, dass sich bessere Testleistungen für Jungen in den Untertests *Wahrnehmung Visuell*, *Gedächtnis Räumlich-Visuell* und *Denken Bildlich* und bessere Testleistungen für Mädchen in den Untertests *Aufmerksamkeit Selektiv*, *Gedächtnis Phonologisch* und *Gedächtnis Auditiv* ergäben. Hinsichtlich Varianzunterschieden erwarten wir, dass sich in den Untertests keine Geschlechtsunterschiede zeigen.

Methoden

Stichprobe

Für diese Studie wurden die Normierungsstichproben der IDS-P und IDS verwendet. Die Normierungsstichprobe der IDS-P umfasst 700 Kinder (50% Jungen) im Alter von 3;0 bis 5;11 Jahren ($M=4.46$, $SD=0.82$); 397 (57%) Kinder waren in der Schweiz, 129 (18%) in Deutschland und 174 (25%) in Österreich wohnhaft. Die Rekrutierung erfolgte über Kindertagesstätten, Spielgruppen und Kindergärten. Die Normierung fand in den Jahren 2010 bis 2012 statt. Die Normierungsstichprobe der IDS umfasst 1330 Kinder (50% Jungen) im Alter von 5;0 bis 10;11 Jahren ($M=7.96$, $SD=1.64$); 744 (56%) waren in der Schweiz, 343 (26%) in Deutschland und 243 (18%) in Österreich wohnhaft. Die Rekrutierung erfolgte über Spielgruppen, Kindergärten und Grundschulen. Die Normierung fand in den Jahren 2007 bis 2008 statt. Die Normierungen beider Testverfahren erfolgten geschlechtsübergreifend, es liegen somit keine separaten Normtabellen für die Geschlechter vor. Weitere Informationen können den Manualen (Grob, Meyer, et al., 2013; Grob, Reimann, et al., 2013) entnommen werden.

Verfahren

Geschlechtsinvarianz und -differenzen in den IDS

Die IDS-P und IDS sind Einzeltestverfahren und erfassen sowohl die Intelligenz als auch die allgemeine Entwicklung in fünf verschiedenen Funktionsbereichen (*Psychomotorik, Sozial-Emotionale Kompetenz, Denken Logisch-Mathematisch, Sprache, Umgang mit der Testsituation* in den IDS-P respektive *Leistungsmotivation* in den IDS). Für die vorliegende Studie wurden die Intelligenz und die dazugehörigen Untertests miteinbezogen.

Die Intelligenzeinschätzung in den IDS-P und IDS erfolgt mit vergleichbaren, altersangepassten Aufgaben in sieben Untertests. Der Untertest *Wahrnehmung Visuell* beinhaltet die Aufgabe, Abbildungen nach ihrer Länge zu ordnen (IDS-P: 8 Items; IDS: 7 Items). *Aufmerksamkeit Selektiv* umfasst die Aufgabe, Stimuli mit einem bestimmten Merkmal aus einer Menge verschiedener Stimuli möglichst schnell und richtig auszusortieren (IDS-P) respektive durchzustreichen (IDS). Dabei werden ein Wert für richtig aussortierte/durchgestrichene Stimuli sowie ein Fehlerwert (fälschlicherweise aussortierte/durchgestrichene und fälschlicherweise ausgelassene Stimuli) errechnet. In *Gedächtnis Phonologisch* müssen Kunstwörter (IDS-P; 15 Items) respektive Buchstaben- und Zahlenfolgen (IDS; 12 Items) unmittelbar und in derselben Reihenfolge wiedergegeben werden, in *Gedächtnis Räumlich-Visuell* müssen geometrische Figuren aus einer Auswahl ähnlicher Figuren wiedererkannt werden (IDS-P: 10 Items; IDS: 11 Items). *Gedächtnis Auditiv* hat zur Aufgabe, eine semantisch sinnvolle Geschichte frei und gestützt zu erinnern (IDS-P: 10 Items; IDS: 24 Items). *Denken Bildlich* erfordert das Nachlegen von geometrischen Figuren mit Hilfe von drei- und viereckigen Klötzchen (IDS-P: 12 Items; IDS: 10 Items), während in *Denken Konzeptuell* bei Bildern das Konzept erkannt und dazu aus einer Auswahl weiterer Bilder das zum Konzept passende Bild ausgewählt werden muss (IDS-P: 11 Items; IDS: 12 Items).

Die Durchführung der sieben Untertests dauert etwa 45 Minuten. Die Wertpunkte der Untertests sind mit $M=10$ und $SD=3$ normiert und können zu einem Wert für die allgemeine

Geschlechtsinvarianz und -differenzen in den IDS

Intelligenz mit $M=100$ und $SD=15$ verrechnet werden. Die Reliabilität für die allgemeine Intelligenz liegt mit Cronbachs $\alpha=.91$ für die IDS-P und $\alpha=.92$ für die IDS im hohen Bereich. Die Reliabilitäten für die Untertests liegen im niedrigen bis hohen Bereich und reichen von $\alpha=.68$ bis $\alpha=.96$.

Statistisches Vorgehen

Die statistischen Analysen erfolgten in mehreren Schritten. In einer Voranalyse wurden die IDS-P und IDS bezüglich ihrer Messinvarianz über das Alter auf Ebene der latenten Untertests überprüft. Bei gegebener skalarer Invarianz auf dieser Ebene wurde die Messinvarianz auf Ebene der latenten allgemeinen Intelligenz getestet. Bei gegebener skalarer Invarianz auf dieser Ebene wurde in einer ersten Hauptanalyse die Messinvarianz über das Geschlecht auf Ebene der latenten Untertests überprüft. Bei gegebener skalarer Invarianz auf dieser Ebene wurde die Messinvarianz auf Ebene der latenten allgemeinen Intelligenz getestet. Sofern skalare Invarianz auf dieser Ebene bestand, wurden als zweite Hauptanalyse manifeste Mittelwert- und Varianzdifferenzen in der allgemeinen Intelligenz und in den dazugehörigen Untertests zwischen den Geschlechtern untersucht.

Für die Voranalyse der Messinvarianz über das Alter wurden für beide Testverfahren in einem ersten Schritt konfirmatorische Faktorenanalysen (CFAs) auf Ebene der Untertests und in einem zweiten Schritt CFAs auf Ebene der allgemeinen Intelligenz durchgeführt. Bei den IDS-P wurden die Altersgruppen 3;0–3;11 ($n=223$), 4;0–4;11 ($n=253$) und 5;0–5;11 ($n=224$) miteinander verglichen. Bei den IDS wurden die Altersgruppen 5;0–6;11 ($n=412$), 7;0–8;11 ($n=492$) und 9;0–10;11 ($n=426$) miteinander verglichen. Für die Überprüfung der Messinvarianz über das Geschlecht wurden für beide Testverfahren in einem ersten Schritt CFAs auf Ebene der Untertests und in einem zweiten Schritt CFAs auf Ebene der allgemeinen Intelligenz durchgeführt. Es wurden jeweils Jungen (IDS-P: $n=350$; IDS: $n=660$) und Mädchen (IDS-P: $n=350$; IDS: $n=670$) miteinander verglichen.

Geschlechtsinvarianz und -differenzen in den IDS

Bei der Modellierung der latenten Untertests wurden die Items als manifeste Indikatoren genommen, als Dateneingabe wurden die Item-Rohwerte verwendet. Bei der Modellierung der latenten allgemeinen Intelligenz wurden die Untertests als manifeste Indikatoren genommen, als Dateneingabe wurden die standardisierten Werte der Untertests verwendet. Als Schätzmethode wurde *Full Information Maximum Likelihood* (FIML) gewählt, die Voraussetzungen (zufällig fehlende Werte, Normalverteilung) waren gegeben. Es wurden für beide Testverfahren pro CFA je fünf Modelle spezifiziert: Modell 1 (Ausgangsmodell) dient der Überprüfung der allgemeinen Faktorenstruktur getrennt für die Gruppen (Einzelgruppenanalyse). Hierzu wurden die Faktorladung des ersten Indikators auf 1 und der y-Achsenabschnitt des ersten Indikators auf 0 fixiert. Darauf aufbauend wurden vier Modelle für Multigruppenvergleiche geschätzt: Modell 2 (konfigurale Invarianz) dient der Testung der allgemeinen Faktorenstruktur. Hierfür wurden bei allen Gruppen die Faktorladung des ersten Indikators auf 1 und der y-Achsenabschnitt des ersten Indikators auf 0 fixiert. Bei Modell 3 (metrische Invarianz) wurden zusätzlich die jeweiligen Faktorladungen der restlichen Indikatoren über die Gruppen hinweg gleichgesetzt. Bei Modell 4 (skalare Invarianz) wurden zusätzlich die jeweiligen y-Achsenabschnitte über die Gruppen hinweg gleichgesetzt. Bei Modell 5 (Invarianz der Fehlervarianzen) erfolgte schließlich eine zusätzliche Gleichsetzung der Fehlervarianzen der Indikatoren über die Gruppen. War vollständige Invarianz bei einem Modell nicht gegeben, erfolgte eine Überprüfung von partieller Invarianz durch eine freie Schätzung von einzelnen Parametern bei gleichzeitiger Gleichsetzung der restlichen Parameter (Byrne et al., 1989)¹. Für die Beurteilung der Modellpassung wurden die deskriptiven Fit-Indizes Comparative Fit Index (CFI), Mc Donalds Non-Centrality Index (NCI), Gamma Hat (GH) und Root Mean Square Error of Approximation (RMSEA) herangezogen. Gemäß Hu und Bentler (1999) und van de Schoot, Lugtig und Hox (2012) kann bei Werten von CFI und GH $\geq .95$, NCI $\geq .90$ und RMSEA $\leq .06$ von einer guten und bei einem

Geschlechtsinvarianz und -differenzen in den IDS

Wert von $CFI \geq .90$ von einer akzeptablen Modellpassung ausgegangen werden. Sofern drei der vier Fit-Indizes die aufgeführten Grenzwerte nicht unter- respektive überschreiten, wurde die Modellpassung als gut respektive akzeptabel bewertet². Aufgrund der großen Stichproben wurden die Chi-Quadrat-Werte nicht interpretiert (Barrett, 2007). Für die Überprüfung der Messinvarianz wurden die Veränderungen im CFI, NCI und GH berücksichtigt. Gemäß Cheung und Rensvold (2002) und Milfont und Fischer (2010) wird bei Werten von ΔCFI und $\Delta GH \leq .01$ beziehungsweise $\Delta NCI \leq .02$ von keiner substantiellen Verschlechterung der Modellpassung ausgegangen. Sofern $\Delta CFI \leq .01$ und einer der beiden anderen Fit-Indizes die aufgeführten Grenzwerte nicht überschreitet, wurde die Modellveränderung als nicht signifikant schlechter befunden³.

Mögliche manifeste Leistungsdifferenzen in der allgemeinen Intelligenz und in den Untertests wurden separat für die beiden Testverfahren basierend auf standardisierten Werten untersucht. Die Mittelwertdifferenzen wurden anhand von *t*-Tests, die Varianzdifferenzen anhand von Levene-Tests berechnet. Des Weiteren wurden Effektstärken *d* nach Cohen (1988) für Mittelwertunterschiede sowie *Variance Ratios* (VRs) nach Feingold (1992) für Varianzunterschiede berechnet. $d \geq .20$ stellt einen kleinen, $d \geq .50$ einen mittleren und $d \geq .80$ einen großen Effekt dar. Ein $VR = 1.00$ deutet auf Varianzhomogenität hin, während ein $VR > 1.00$ eine größere Varianz für Jungen und ein $VR < 1.00$ eine größere Varianz für Mädchen bedeutet. VRs zwischen 1.10 und 0.90 gelten als vernachlässigbar. Für die Untertests wurde das Alpha-Niveau mit Hilfe der Korrektur nach Hommel (1988) adjustiert, um der Alpha-Fehler-Inflation entgegenzuwirken, welche durch multiples Testen in derselben Stichprobe entstehen kann. Geschlechtsdifferenzen wurden als bedeutsam interpretiert, sofern sowohl Signifikanz ($p < .05$) als auch mindestens kleine Effekte ($d \geq .20$) respektive deutlich ungleiche VRs ($VR > 1.10$, $VR < 0.90$) erreicht wurden. Eine Poweranalyse mit G*Power (Faul, Erdfelder, Lang & Buchner, 2007) ergab, dass mit den untersuchten Stichproben klei-

Geschlechtsinvarianz und -differenzen in den IDS

ne Effekte (d) mit $\alpha=.05$ und Power=.80 detektiert werden können. Die Analysen wurden mit den Statistikprogrammen AMOS und SPSS (IBM, 2012) vorgenommen.

Ergebnisse

Ein Überblick über die deskriptive Statistik mit Reliabilitäten, Mittelwerten, Standardabweichungen und Korrelationen für die allgemeine Intelligenz und die Untertests für die IDS-P und IDS ist in Tabelle 1 dargestellt.

Tabelle 1

Voranalysen: Messinvarianz über das Alter

In Bezug auf das Alter zeigten die Untertests einen akzeptablen bis guten Fit des Ausgangsmodells für die einzelnen Altersgruppen und weiter konfigurale, metrische und vollständig oder partiell skalare Invarianz sowie vollständige oder partielle Invarianz der Fehlervarianzen mit einem guten bis akzeptablen Fit. Die allgemeine Intelligenz zeigte einen akzeptablen bis guten Fit des Ausgangsmodells für die einzelnen Altersgruppen und weiter konfigurale, metrische und skalare Invarianz sowie Invarianz der Fehlervarianzen. Die Ergebnisse sind als Elektronisches Supplement ESM 1 (Tabelle 1: Untertests IDS-P; Tabelle 2: Untertests IDS; Tabelle 3: allgemeine Intelligenz IDS-P und IDS) verfügbar.

Hauptanalysen 1: Messinvarianz über das Geschlecht

In einem ersten Teil wurden die Verfahren auf Messinvarianz über das Geschlecht untersucht. Die Ergebnisse sind als ESM 2 (Tabelle 1: Untertests IDS-P; Tabelle 2: Untertests IDS; Tabelle 3: allgemeine Intelligenz IDS-P und IDS) verfügbar.

Bei den Untertests zeigten die Ausgangsmodelle für beide Testverfahren und beide Geschlechter einen akzeptablen bis guten Fit. Bei den IDS-P zeigten die Multigruppenvergleiche für die Untertests *Wahrnehmung Visuell*, *Gedächtnis Räumlich-Visuell*, *Gedächtnis*

Geschlechtsinvarianz und -differenzen in den IDS

Auditiv, *Denken Bildlich* und *Denken Konzeptuell* konfigurale, metrische und skalare Invarianz sowie Invarianz der Fehlervarianzen. Der Untertest *Gedächtnis Phonologisch* wies konfigurale, partiell metrische und partiell skalare Invarianz sowie partielle Invarianz der Fehlervarianzen auf. Der Untertest *Aufmerksamkeit Selektiv* zeigte metrische und skalare Invarianz sowie Invarianz der Fehlervarianzen.

Bei den IDS zeigte der Untertest *Gedächtnis Auditiv* konfigurale, metrische und skalare Invarianz sowie Invarianz der Fehlervarianzen und der Untertest *Denken Bildlich* konfigurale, metrische und skalare Invarianz sowie partielle Invarianz der Fehlervarianzen. Die Untertests *Wahrnehmung Visuell*, *Gedächtnis Phonologisch*, *Gedächtnis Räumlich-Visuell* und *Denken Konzeptuell* wiesen konfigurale, partiell metrische und partiell skalare Invarianz sowie partielle Invarianz der Fehlervarianzen auf. Der Untertest *Aufmerksamkeit Selektiv* zeigte metrische und skalare Invarianz sowie Invarianz der Fehlervarianzen.

Bei der allgemeinen Intelligenz zeigten die Ausgangsmodelle für beide Testverfahren und beide Geschlechter einen akzeptablen bis guten Fit. Bei den IDS-P zeigten die Multigruppenvergleiche konfigurale und metrische Invarianz. Skalare Invarianz musste zurückgewiesen werden, es wurde jedoch partiell skalare Invarianz und, basierend auf diesem Modell, partielle Invarianz der Fehlervarianzen erreicht. Bei den IDS zeigten die Multigruppenvergleiche konfigurale Invarianz. Metrische Invarianz musste zurückgewiesen werden, es wurde jedoch partiell metrische Invarianz und, basierend auf diesem Modell, partiell skalare Invarianz und partielle Invarianz der Fehlervarianzen erreicht.

Hauptanalysen 2: Mittelwert- und Varianzdifferenzen

In einem zweiten Teil wurden manifeste Mittelwert- und Varianzdifferenzen in der allgemeinen Intelligenz sowie in den Untertests untersucht. Die Ergebnisse sind in Tabelle 2 dargestellt. Beide Testverfahren zeigten keine Unterschiede in der allgemeinen Intelligenz im Mittelwert und in der Varianz. In den IDS-P zeigte sich eine signifikante und kleine Mit-

Geschlechtsinvarianz und -differenzen in den IDS

telwertdifferenz zu Gunsten von Jungen im Untertest *Denken Bildlich* mit $\Delta M=0.68$ ($p_{\text{kor}}=.02$, $d=0.23$), während Mädchen höhere Mittelwerte in den Untertests *Aufmerksamkeit Selektiv* mit $\Delta M=-0.95$ ($p_{\text{kor}}<.001$, $d=-0.32$) und *Gedächtnis Auditiv* mit $\Delta M=-0.70$ ($p_{\text{kor}}=.02$, $d=-0.24$) aufwiesen. In den anderen IDS-P-Untertests fanden sich keine Mittelwertunterschiede. In den IDS zeigte sich eine signifikante und kleine Mittelwertdifferenz zu Gunsten von Jungen im Untertest *Denken Bildlich* mit $\Delta M=1.29$ ($p_{\text{kor}}<.001$, $d=0.45$), in den anderen IDS-Untertests fanden sich keine Mittelwertunterschiede. Sowohl in den IDS-P wie auch in den IDS unterschieden sich die Varianzen von Jungen und Mädchen in keinem der Untertests bedeutsam.

Tabelle 2

Diskussion

Das Ziel der vorliegenden Studie war, die Messinvarianz über das Alter und über das Geschlecht auf Ebene der Untertests und auf Ebene der allgemeinen Intelligenz sowie Leistungsdifferenzen zwischen den Geschlechtern in der allgemeinen Intelligenz und in den Untertests der IDS-P und IDS zu untersuchen. Die Analysen bezüglich des Alters zeigten jeweils vollständige oder partielle Invarianz der Fehlervarianzen für die Untertests und vollständige Invarianz der Fehlervarianzen für die allgemeine Intelligenz. Bei der Überprüfung der Messinvarianz hinsichtlich Geschlecht resultierte bei den Untertests für die IDS-P für sechs Untertests vollständige und für einen Untertest partielle Invarianz der Fehlervarianzen. Bei den IDS ergab sich für zwei Untertests vollständige und für die restlichen Untertests partielle Invarianz der Fehlervarianzen. Bei der allgemeinen Intelligenz wiesen beide Verfahren partielle Invarianz der Fehlervarianzen auf.

Dies bedeutet zum einen, dass die von den Autoren postulierte Struktur der beiden Verfahren (Grob, Meyer, et al., 2013; Grob, Reimann, et al., 2013) für die verschiedenen

Geschlechtsinvarianz und -differenzen in den IDS

Altersgruppen und für beide Geschlechter bestätigt wurde und demzufolge beide Testverfahren das gleiche Konstrukt äquivalent messen (Millsap & Kwok, 2004). Zum anderen erweitern diese Ergebnisse die Konstruktvalidierung der beiden Verfahren im Sinne eines Nachweises für die faktorielle Validität getrennt nach Alter und Geschlecht (Moosbrugger & Kelava, 2008).

Weiter wurden manifeste Leistungsunterschiede im Hinblick auf Mittelwerte und Varianzen in der allgemeinen Intelligenz und in den Untertests untersucht. In der allgemeinen Intelligenz wurden keine Mittelwertdifferenzen erwartet, was bestätigt werden konnte. Dieser Befund geht mit der *Geschlechtsähnlichkeitshypothese* (Hyde, 2005) sowie mit der Entwicklungstheorie von Lynn (1994) und Ergebnissen aus anderen Normierungsstudien (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Savage-McGlynn, 2012; Sellers et al., 2002) einher.

Für die allgemeine Intelligenz wurden größere Varianzen bei Jungen erwartet, was weder für die IDS-P noch für die IDS bestätigt werden konnte. Diese Resultate decken sich mit Befunden anderer Normierungsstudien, die ebenfalls keine Varianzunterschiede zwischen den Geschlechtern aufzeigen konnten (z.B. Savage-McGlynn, 2012), widersprechen aber der *Hypothese der größeren männlichen Variabilität* (Hyde & Mertz, 2009) und den Ergebnissen von Arden und Plomin (2006) und Deary et al. (2003). Ein möglicher Grund für die hier gefundenen Ergebnisse kann in der Zusammensetzung der Normierungsstichproben liegen: Die Rekrutierung der Probanden erfolgte über Kindertagesstätten, Spielgruppen, Kindergärten und Grundschulen, während Sonderschulen nicht spezifisch miteinbezogen wurden (Grob, Meyer, et al., 2013). Dies könnte zur Folge gehabt haben, dass insbesondere Kinder mit einer Intelligenzminderung sowie hochbegabte Kinder, welche gegebenenfalls den Unterricht in Privatschulen besuchen, in den vorliegenden Stichproben unterrepräsentiert waren.

Geschlechtsinvarianz und -differenzen in den IDS

Bezüglich spezifischen Intelligenzaspekten erwarteten wir bessere Leistungen für Jungen in den Untertests *Wahrnehmung Visuell*, *Gedächtnis Räumlich-Visuell* und *Denken Bildlich*. In *Denken Bildlich* zeigten Jungen höhere Mittelwerte, in den anderen beiden Untertests fanden sich keine Unterschiede. Der Geschlechtsunterschied in *Denken Bildlich* entspricht der *Differenzhypothese* (Maccoby & Jacklin, 1974) und weiteren Normierungsstudien zu visuell-räumlichen Fähigkeiten (Reynolds et al., 2008), während die Resultate der anderen Untertests im Widerspruch dazu stehen. Es existieren jedoch auch Normierungsstudien, welche ebenfalls keine Unterschiede in visuell-räumlichen Fähigkeiten aufzeigen konnten (Camarata & Woodcock, 2006). Ein Grund für die hier gefundenen Ergebnisse könnte sein, dass es sich beim Untertest *Wahrnehmung Visuell*, in welchem Abbildungen nach ihrer Länge geordnet werden müssen, und *Gedächtnis Räumlich-Visuell*, in welchem geometrische Figuren aus einer Auswahl ähnlicher Figuren wiedererkannt werden müssen, eher um basale visuell-räumliche Aufgaben handelt und möglicherweise Geschlechtsunterschiede in solchen nicht zum Vorschein kommen. *Denken Bildlich* hingegen erfordert das Nachlegen von geometrischen Figuren mit Hilfe von Klötzchen und somit komplexere visuell-räumliche Fähigkeiten wie räumliches Vorstellungsvermögen und mentale Rotation, wie dies auch in Aufgabentypen wie beispielsweise *Musterkonstruktion* und *Mustererinnerung* in den *Differential Ability Scales–Second Edition* (DAS-II; Elliott, 2007) der Fall ist. In derartigen Aufgaben und ähnlichen Untertests konnten Normierungsstudien Unterschiede (Reynolds et al., 2008) oder zumindest Tendenzen (Keith et al., 2008; Keith et al., 2011) zu Gunsten von Jungen finden.

Weiter erwarteten wir bessere Leistungen für Mädchen in den Untertests *Aufmerksamkeit Selektiv*, *Gedächtnis Phonologisch* und *Gedächtnis Auditiv*. In *Aufmerksamkeit Selektiv* muss eine Zielfigur aus einer Menge ähnlicher Figuren möglichst schnell und richtig aussortiert respektive durchgestrichen werden. In den IDS-P erzielten Mädchen einen höhe-

Geschlechtsinvarianz und -differenzen in den IDS

ren Mittelwert als Jungen, was anderen Normierungsstudien entspricht (Camarata & Woodcock, 2006; Goldbeck et al., 2010; Keith et al., 2008; Keith et al., 2011). Dieser Geschlechtsunterschied war in den IDS nicht ersichtlich. Weiter wies der Untertest *Gedächtnis Phonologisch*, welcher die Wiedergabe von Kunstwörtern respektive Zahlen- und Buchstabenfolgen erfordert, keine Unterschiede auf. Dies widerspricht Normierungsstudien, welche bessere Leistungen zu Gunsten von Mädchen im Kurzzeitgedächtnis aufzeigen konnten (Keith et al., 2011), wobei gewisse Studien ebenfalls keine Unterschiede in diesen Bereichen finden konnten (Camarata & Woodcock, 2006; Keith et al., 2008). Beispielsweise stellen die WISC-IV-Untertests *Zahlen nachsprechen* und *Buchstaben-Zahlen-Folgen* des Index *Arbeitsgedächtnis* (Petermann & Petermann, 2011) zu IDS-P und IDS *Gedächtnis Phonologisch* vergleichbare Untertests dar, in welchen sich in der Normierungsstichprobe ebenfalls keine Geschlechtsunterschiede zeigten (Goldbeck et al., 2010). Ferner resultierten in *Gedächtnis Auditiv* (Erinnerung einer semantisch sinnvollen Geschichte) in den IDS-P höhere Testwerte für Mädchen. Dieses Ergebnis entspricht anderen Normierungsstudien, die aufzeigen konnten, dass Mädchen einen Leistungsvorsprung im Langzeitgedächtnis aufweisen (Keith et al., 2011). Da dieser Untertest sprachlastig ist, geht das Resultat außerdem mit der *Differenzhypothese* (Maccoby & Jacklin, 1974) konform. Dieser Geschlechtsunterschied war in den IDS nicht ersichtlich.

Im Untertest *Denken Konzeptuell* muss bei Bildern das Konzept erkannt und dazu aus einer Auswahl weiterer Bilder das zum Konzept passende Bild ausgewählt werden. Hier waren keine Unterschiede zwischen Jungen und Mädchen feststellbar. Schließlich war die Varianz der Geschlechter in allen Untertests in beiden Testverfahren vergleichbar, was mit den Befunden von Palejwala und Fine (2015) und Reynolds et al. (2008) konform ist.

Eine Stärke der vorliegenden Studie stellt die Untersuchung von Normierungsstichproben dar. Die IDS-P und IDS verfügen über große Stichproben, die repräsentativ hinsicht-

Geschlechtsinvarianz und -differenzen in den IDS

lich Alter, Geschlecht und sozio-ökonomischem Status für die Population sind. Es ist positiv hervorzuheben, dass eine vorgängige Überprüfung der Messinvarianz über das Alter und über das Geschlecht erfolgte. Es ist kritisch anzumerken, dass die Normierungsstichproben nicht explizit Kinder mit Intelligenzminderung oder Hochbegabung beinhalten. Weitere Forschung ist notwendig, um zu klären, ob die hier gefundenen Ergebnisse auf diese Kinder generalisiert werden können. Zukünftige Studien sollten Geschlechtsdifferenzen in verschiedenen Altersgruppen untersuchen, da sich die Intelligenz mit dem Alter ausdifferenziert (vgl. Rost, 2009). Für die weiterführende Validierung der beiden Testverfahren könnten außerdem Alters- und Geschlechtsdifferenzen in Adaptionen der IDS-P und IDS in anderen Sprachgruppen untersucht werden. Die IDS-P wurden in Polen (Fecenec, Jaworowska & Matczak, 2015), die IDS in Polen (Jaworowska, Matczak & Fecenec, 2011), Tschechien (Krejčířová, Urbánek, Širůček & Jabůrek, 2013) und Italien (Ferri, Rea & Casagrande, 2015) adaptiert, was zudem eine länderübergreifende Untersuchung der Faktorenstruktur ermöglichen würde.

Zusammenfassend lässt sich festhalten, dass die IDS-P und IDS invariant bezüglich ihrer Struktur über das Alter und das Geschlecht sind und folglich faktorielle Validität aufweisen. Die Leistungen in der allgemeinen Intelligenz sind zwischen den Geschlechtern vergleichbar. In einigen Untertests werden hingegen kleine geschlechtsspezifische Leistungsunterschiede sichtbar: Jungen zeigen in beiden Verfahren Stärken in *Denken Bildlich*, Mädchen in *Aufmerksamkeit Selektiv* und *Gedächtnis Auditiv* in den IDS-P. Insgesamt kann basierend auf den vorliegenden Ergebnissen gefolgert werden, dass die IDS-P und IDS eine faire Einschätzung der Intelligenz für beide Geschlechter trotz geschlechtsunspezifischer Normtabellen in der psychologischen Praxis ermöglichen.

Elektronische Supplemente (ESM)

ESM1. Invarianzanalysen über das Alter (ESM1.pdf)

Geschlechtsinvarianz und -differenzen in den IDS

ESM2. Invarianzanalysen über das Geschlecht (ESM2.pdf)

Literatur

- Arden, R. & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, 41, 39-48.
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42, 815-824.
- Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191-244.
- Byrne, B.M., Shavelson, R.J. & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.
- Camarata, S. & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, 34, 231-252.
- Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.
- Cattell, R.B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, 38, 592.
- Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NY: Erlbaum.
- Cramer, J.S. (2012). Childhood intelligence and adult mortality, and the role of socio-economic status. *Tinbergen Institute Discussion Paper*, 12, 1-22.
- Deary, I.J., Thorpe, G., Wilson, V., Starr, J.M. & Whalley, L.J. (2003). Population sex differences in IQ at age 11: The Scottish Mental Survey 1932. *Intelligence*, 31, 533-542.
- Deary, I.J., Whalley, L.J., Lemmon, H., Crawford, J.R. & Starr, J.M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence*, 28, 49-55.
- Elliott, C.D. (2007). *Differential Ability Scales—Second Edition (DAS-II)*. San Antonio, TX: Harcourt.

Geschlechtsinvarianz und -differenzen in den IDS

- Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.
- Fecenec, D., Jaworowska, A. & Matczak, A. (2015). *IDS-Skale Inteligencji i Rozwoju dla Dzieci w Wiekui Przedszkolnym*. Warschau: Hogrefe.
- Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62, 61-84.
- Ferri, R., Rea, M. & Casagrande, M. (2015). *IDS-Intelligence and Development Scales*. Florenz: Hogrefe.
- Goldbeck, L., Daseking, M., Hellwig-Brida, S., Waldmann, H.C. & Petermann, F. (2010). Sex differences on the German Wechsler Intelligence Test for Children (WISC-IV). *Journal of Individual Differences*, 31, 22-28.
- Goldstein, S., Princiotta, D. & Naglieri, J.A. (2015). *Handbook of intelligence. Evolutionary theory, historical perspective, and current concepts*. New York, NY: Springer.
- Grob, A., Meyer, C.S. & Hagmann-von Arx, P. (2013). *Intelligence and Development Scales (IDS)*. 2., überarb. Aufl. Bern: Huber.
- Grob, A., Reimann, G., Gut, J. & Frischknecht, M.-C. (2013). *Intelligence and Development Scales-Preschool (IDS-P)*. Bern: Huber.
- Gut, J., Reimann, G. & Grob, A. (2012). Kognitive, sprachliche, mathematische und sozial-emotionale Kompetenzen als Prädiktoren späterer schulischer Leistungen: Können die Leistungen eines Kindes in den IDS dessen Schulnoten drei Jahre später vorhersagen? *Zeitschrift für Pädagogische Psychologie*, 26, 213-220.
- Gygi, J.T., Hagmann-von Arx, P., Schweizer, F. & Grob, A. (2017). The predictive validity of four intelligence tests for school grades: A small sample longitudinal study. *Frontiers in Psychology*, 8: 375.
- Hagmann-von Arx, P., Grob, A., Petermann, F. & Daseking, M. (2012). Konkurrente Validität des HAWIK-IV und der Intelligence and Development Scales (IDS). *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 40, 41-50.
- Hagmann-von Arx, P., Meyer, C.S. & Grob, A. (2008). Assessing intellectual giftedness with the WISC-IV and the IDS. *Zeitschrift für Psychologie*, 216, 172-179.
- Hagmann-von Arx, P., Petermann, F. & Grob, A. (2013). Konvergente und diskriminante Validität der WISC-IV und der Intelligence and Development Scales (IDS) bei Kindern mit Migrationshintergrund. *Diagnostica*, 29, 170-182.

Geschlechtsinvarianz und -differenzen in den IDS

- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383-386.
- Horn, J.L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities*. University of Illinois, Champaign: Unpublished Doctoral Dissertation.
- Hu, L.T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.
- Hyde, J.S. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581-592.
- Hyde, J.S. & Mertz, J.E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, 106, 8801-8807.
- IBM. (2012). *SPSS Statistics for Mac*. Armonk, NY: IBM Corp.
- Jaworowska, A., Matczak, A. & Fecenec, D. (2011). *IDS-Skale Inteligencji i Rozwoju dla Dzieci w wieku 5-10 lat*. Warschau: Hogrefe.
- Johnson, W. & Bouchard, T.J.J. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence*, 35, 23-39.
- Keith, T.Z., Reynolds, M.R., Patel, P.G. & Ridley, K.P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock-Johnson III Tests of Cognitive Abilities. *Intelligence*, 36, 502-525.
- Keith, T.Z., Reynolds, M.R., Roberts, L.G., Winter, A.L. & Austin, C.A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales-Second Edition. *Intelligence*, 39, 389-404.
- Kramer, J. (1972). *Intelligenztest: Mit einer Einführung in Theorie und Praxis der Intelligenzprüfung*. 4., rev. Aufl. Solothurn: Antonius.
- Krejčířová, D., Urbánek, T., Širůček, J. & Jabůrek, M. (2013). *IDS-Intelligenční a vývojová škála pro děti ve věku 5-10 let Prague*. Prag: Hogrefe.
- Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, 17, 257-271.
- Maccoby, E.E. & Jacklin, C.N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- McGrew, K.S. (2005). The Cattell-Horn-Carroll Theory of cognitive abilities: Past, present, and future. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary*

Geschlechtsinvarianz und -differenzen in den IDS

- intellectual assessment: Theories, tests, and issues (2nd ed.)* (pp. 136-182). New York, NY: Guilford Press.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- Meyer, C.S., Hagmann-von Arx, P. & Grob, A. (2009). Die Intelligence and Development Scale Sozial-Emotionale Kompetenz (IDS-SEK). *Diagnostica*, 55, 234-244.
- Meyer, C.S., Hagmann-von Arx, P., Lemola, S. & Grob, A. (2010). Correspondence between the general ability to discriminate visual stimuli and general intelligence. *Journal of Individual Differences*, 31, 46-56.
- Milfont, T.L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural-research. *International Journal of Psychological Research*, 3, 111-121.
- Millsap, R.E. & Kwok, O.M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93-115.
- Moosbrugger, H. & Kelava, A. (2008). *Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien)*. Berlin: Springer.
- Palejwala, M.H. & Fine, J.G. (2015). Gender differences in latent cognitive abilities in children aged 2 to 7. *Intelligence*, 48, 96-108.
- Petermann, F. & Petermann, U. (2011). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) – Deutsche Version*. Frankfurt a. M.: Pearson Assessment.
- Reynolds, M.R., Keith, T.Z., Ridley, K.P. & Patel, P.G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, 36, 236-260.
- Rost, D.H. (2009). *Intelligenz: Fakten und Mythen* (korr. Nachdr.). Weinheim: Beltz.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F. & Spinath, F.M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137.
- Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences*, 53, 137-141.
- Schaie, K.W. (1994). The course of adult intellectual development. *American Psychologist*, 49, 304-313.

Geschlechtsinvarianz und -differenzen in den IDS

- Sellers, A.H., Burns, W.J. & Guyrke, J. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence-Revised as a function of stratification variables. *Applied Neuropsychology*, 9, 65-73.
- Spearman, C. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology*, 15, 201-292.
- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, 35, 401-426.
- van de Schoot, R., Lugtig, P. & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9, 486-492.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV)*. San Antonio, TX: The Psychological Corporation.
- Widaman, K.F. & Reise, S.P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K.J. Bryant, M. Windle & S.G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.
- Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U. & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, 33, 292-296.

¹ Da der Untertest *Aufmerksamkeit Selektiv* nur aus 2 Werten besteht, wurden von Beginn an beide Faktorladungen auf 1 und der y-Achsenabschnitt des ersten Indikators auf 0 fixiert (Modell 1 und Modell 3). Die Fixierung beider Ladungen auf 1 impliziert, dass dieses Modell metrische Invarianz aufweist und somit Modell 3 entspricht. Bei Modell 4 wurde zusätzlich der y-Achsenabschnitt des zweiten Indikators gleichgesetzt; bei Modell 5 wurden zusätzlich die Fehlervarianzen der Indikatoren gleichgesetzt.

² *Aufmerksamkeit Selektiv*: Modell 1 und 3: keine Bewertung.

³ *Aufmerksamkeit Selektiv*: Modell 4 vs. 3: keine Bewertung.

Geschlechtsinvarianz und -differenzen in den IDS

Tabelle 1

Deskriptive Statistik für die IDS-P und IDS

Stich- probe	Vari- ablen	Cronbachs Alpha	M(SD)			r ²							
			Gesamt	Jungen	Mädchen	IQ	WV	AS	GP	GR	GA	DB	DK
IDS-P	IQ	.91	99.94 (15.03)	98.39 (15.44)	100.68 (14.55)	–	.60***	.43***	.50***	.61***	.49***	.61***	.58***
	WV	.84	10.04 (2.99)	9.93 (3.07)	10.16 (2.91)	.60***	–	.14**	.16**	.35***	.12*	.38***	.20***
	AS	.93	9.98 (3.00)	9.50 (2.90)	10.45 (3.01)	.55***	.25***	–	.06	.15**	.05	.20***	.05
	GP	.72	10.01 (3.02)	9.84 (3.08)	10.19 (2.95)	.45***	.15**	.07	–	.09	.21***	.17**	.18***
	GR	.83	9.95 (2.99)	9.93 (2.99)	9.98 (2.99)	.62***	.22***	.24***	.17**	–	.18**	.26***	.31***
	GA	.81	10.00 (2.98)	9.64 (3.02)	10.34 (2.90)	.54***	.14*	.11	.25***	.23***	–	.13*	.19**
	DB	.85	9.95 (3.03)	10.29 (3.17)	9.61 (2.85)	.61***	.33***	.25***	.04	.30***	.14**	–	.21***
	DK	.81	9.91 (3.05)	9.70 (2.96)	10.11 (3.13)	.59***	.22***	.24***	.06	.28***	.18**	.29***	–
IDS	IQ	.92	100.00 (15.00)	100.02 (14.84)	99.98 (15.17)	–	.54***	.40***	.54***	.58***	.56***	.53***	.60***
	WV	.77	10.02 (2.99)	10.01 (2.99)	10.04 (3.00)	.57***	–	.10*	.18***	.20***	.18***	.27***	.20***
	AS	.96	10.01 (2.96)	9.90 (2.98)	10.11 (2.93)	.50***	.19***	–	.16***	.10*	.15***	.05	.04
	GP	.73	9.98 (3.01)	9.95 (2.91)	10.01 (3.11)	.56***	.24***	.14**	–	.18***	.24***	.11**	.18***
	GR	.68	10.01 (2.98)	9.91 (2.87)	10.10 (3.08)	.55***	.17***	.14**	.16***	–	.20***	.21***	.31***
	GA	.87	9.99 (2.95)	9.77 (2.91)	10.20 (2.99)	.58***	.22***	.18***	.27***	.24***	–	.16***	.30***
	DB	.70	9.93 (2.96)	10.58 (2.86)	9.29 (2.92)	.54***	.24***	.09*	.21***	.22***	.20***	–	.26***
	DK	.75	10.01 (3.01)	9.73 (2.97)	10.29 (3.03)	.52***	.16***	.11**	.15***	.17***	.16***	.20***	–

Geschlechtsinvarianz und -differenzen in den IDS

Anmerkungen. $N_{IDS-P}=627-700$; $N_{IDS}=1203-1330$; IQ=Allgemeine Intelligenz; WV=Wahrnehmung Visuell; AS=Aufmerksamkeit Selektiv; GP=Gedächtnis Phonologisch; GR=Gedächtnis Räumlich-Visuell; GA=Gedächtnis Auditiv; DB=Denken Bildlich; DK=Denken Konzeptuell; Allgemeine Intelligenz: $M=100$, $SD=15$; Untertests: $M=10$, $SD=3$.

^aWerte für Jungen unterhalb und Werte für Mädchen oberhalb der Diagonalen.

* $p<.05$; ** $p<.01$; *** $p<.001$.

Geschlechtsinvarianz und -differenzen in den IDS

Tabelle 2

Manifeste Mittelwert- und Varianzdifferenzen zwischen Jungen und Mädchen in der allgemeinen Intelligenz und in den Untertests in den IDS-P und IDS

Stichprobe	Variable	Mittelwert-differenz	t	p ^a	F	Levene-Test	Effektstärke
		ΔM	t	p ^a	F	p ^a	d
IDS-P	IQ	-2.29	-1.93	.05	0.74	.39	-0.15
	WV	-0.23	-1.00	.96	0.35	1.00	-0.08
	AS	-0.95	-4.19	<.001	0.08	1.00	-0.32
	GP	-0.35	-1.45	.54	0.49	1.00	-0.12
	GR	-0.05	-0.22	1.00	0.03	1.00	-0.02
	GA	-0.70	-3.00	.02	1.14	1.00	-0.24
	DB	0.68	2.92	.02	5.99	.27	0.23
IDS	DK	-0.41	-1.70	.41	0.61	1.00	-0.13
	IQ	0.04	0.06	.96	1.26	.26	0.00
	WV	-0.03	-0.18	1.00	0.03	1.00	-0.01
	AS	-0.21	-1.23	.87	0.04	1.00	-0.07
	GP	-0.06	-0.33	1.00	1.97	1.00	-0.02
	GR	-0.19	-1.18	.87	2.84	1.00	-0.06
	GA	-0.43	-2.64	.05	0.10	1.00	-0.15
	DB	1.29	7.93	<.001	1.86	1.00	0.45
	DK	-0.56	-3.35	<.01	0.41	1.00	-0.19

Anmerkungen. $N_{IDS-P}=627-700$; $N_{IDS}=1203-1330$; IQ=Allgemeine Intelligenz; WV=Wahrnehmung Visuell; AS=Aufmerksamkeit Selektiv; GP=Gedächtnis Phonologisch; GR=Gedächtnis Räumlich-Visuell; GA=Gedächtnis Auditiv; DB=Denken Bildlich; DK=Denken Konzeptuell; Mittelwertdifferenz ΔM: positive Werte=höhere Werte für Jungen, negative Werte=höhere Werte für Mädchen; p^a: bei allgemeiner Intelligenz p, bei Untertests p_{corr} mit Alpha-Korrektur nach Hommel; d=Effektstärke nach Cohen; VR=Variance Ratio nach Feingold; Allgemeine Intelligenz: M=100, SD=15; Untertests: M=10, SD=3.

ESM 1: Messinvarianzanalysen über das Alter

Tabelle 1. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Alter in den Untertests der IDS-P

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	$\Delta\chi^2$	ΔCFI	ΔGH	ΔNCI
1	Ausgangsmodell 3;0-3;11/ 4;0-4;11/ 5;0-5;11	WV	20/20/20	81.81***/57.72***/56.99***	.90/.92/.90	.98/.98/.98	.87/.93/.92	.06/.06/.06				
		AS	0/0/0									
		GP	90/90/90	92.88***/130.26***/72.35***	.91/.91/.91	.97/.96/.96	.99/.92/.100	.05/.06/.06				
		GR	35/35/35	82.93***/45.34***/46.34***	.90/.90/.96	.98/.98/.99	.90/.98/.97	.06/.06/.04				
		GA	35/35/35	70.34***/60.49***/82.69***	.90/.94/.90	.98/.98/.98	.92/.95/.90	.06/.05/.06				
		DB	54/54/54	72.52***/46.56***/84.59***	.91/.91/.90	.98/.97/.97	.96/.100/.93	.05/.06/.06				
2	Konfigurale Invarianz	DK	44/44/44	94.71***/30.83/79.74***	.90/.100/.90	.97/.100/.97	.89/.100/.92	.06/.00/.06				
		WV	60	196.53***	.90	.96	.91	.05				
		AS										
		GP	270	317.07***	.90	.97	.97	.03				
		GR	105	222.41***	.90	.97	.92	.04				
		GA	105	213.53***	.92	.97	.93	.04				
3	Metrische Invarianz	DB	162	216.59***	.92	.94	.96	.05				
		DK	132	225.10***	.92	.98	.94	.03				
		WV	74	237.09***	.90	.96	.89	.05				
		AS	0									
		GP	298	469.45***	.90	.97	.88	.03				
		GR	123	270.04***	.90	.96	.90	.04				
		GA	123	224.95***	.91	.98	.93	.03				
		DB	184	357.83***	.91	.95	.88	.04				
		DK	152	285.95***	.91	.98	.91	.03				
		WV							14	40.56*	<.001	<.001
		AS							28	152.38***	<.001	<.001
		GP							18	47.63*	<.001	.02
		GR							18	11.42	<.001	<.001
		GA							22	141.24***	.01	.01
		DB							20	60.85*	<.001	<.001
		DK										.03

3 vs. 2

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	ΔGH	ANCI
4	Skalare Invarianz	WV	88	268.50***	.90	.95	.88	.05					
		AS	3	16.01	1.00	.99	.99	.06					
		GP	326	613.00***	.90	.96	.81	.03					
		GR	141	353.82***	.90	.96	.86	.04					
		GA	141	250.31***	.90	.98	.92	.03					
		DB	206	565.43***	.90	.95	.77	.04					
		DK	172	328.19***	.90	.97	.89	.03					
		WV							14	31.41*	<.001	.01	.01
		AS							3				
		GP							28	143.55***	<.001	.01	.07
		GR							18	83.78*	<.001	<.001	.04
		GA							18	25.36	.01	<.001	.01
		DB							22	207.60***	.01	<.001	.11
		DK							20	42.24*	.01	.01	.02
5a	Invarianz der Fehlervarianzen	WV	104	758.42***	.84	.83	.63	.09					
		AS	7	17.92	1.00	.98	.99	.06					
		GP	356	801.25***	.82	.72	.73	.09					
		GR	161	979.95***	.82	.83	.56	.08					
		GA	161	254.00***	.89	.95	.94	.04					
		DB	230	922.87***	.87	.80	.61	.08					
		DK	194	706.47***	.85	.82	.69	.08					
		WV							16	489.92***	.06	.12	.25
		AS							4	1.91	<.001	.01	<.001
		GP							30	188.25***	.08	.24	.08
		GR							20	626.13***	.08	.13	.30
		GA							20	3.69	.01	.03	<.001
		DB							24	357.44***	.03	.15	.16
		DK							22	378.28***	.05	.15	.20
5b	Partielle Invarianz der Fehlervarianzen	WV ^b	102	633.66***	.90	.96	.68	.04					
		AS											
		GP ^c	354	792.94***	.90	.96	.73	.03					
		GR ^d	157	650.09***	.90	.95	.70	.04					
		GA											
		DB ^e	226	735.13***	.90	.97	.69	.03					
		DK ^f	192	505.64***	.90	.97	.80	.03					
		WV							14	356.16***	<.001	<.001	.20
		AS							28	179.94***	<.001	<.001	.08
		GP							16	296.27***	<.001	.01	.16
		GR							20	169.70***	<.001	<.001	.08
		GA							20	177.45***	<.001	<.001	.09
		DB											
		DK											

Anmerkungen. $N_{IDS,1} = 627-700$; WV = Wahrnehmung Visuell; AS = Aufmerksamkeit Selektiv; GP = Gedächtnis Phonologisch; GR = Gedächtnis Räumlich-Visuell; GA = Gedächtnis Auditiv; DB = Denken Bildlich; DK = Denken Konzeptuell; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = Mc Donalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^aModell 1: Einzelgruppenanalysen für Altersgruppen; Modelle 2-5: Multigruppenvergleiche zwischen Altersgruppen. Bei AS entspricht Modell 1 metrischer Invarianz.

^bFreie Schätzung der Fehlervarianz von Item 1.

^cFreie Schätzung der Fehlervarianz von Item 12.

^dFreie Schätzung der Fehlervarianzen von Item 1 und 7.

^eFreie Schätzung der Fehlervarianzen von Item 1 und 8.

^fFreie Schätzung der Fehlervarianz von Item 5.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Tabelle 2. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Alter in den Untertests der IDS

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Adf	$\Delta\chi^2$	ACFI	AGH	ANCI
1	Ausgangsmodell 5:0-6;11/ 7:0-8;11/ 9:0-10;11	WV	14/14/14	33.96**/43.59***/34.69**	.93/.93/.90	.99/.99/.99	.98/.97/.98	.06/.06/.06					
		AS	0/0/0										
		GP	54/54/54	97.99***/104.25***/146.40***	.90/.90/.90	.97/.97/.97	.95/.95/.90	.06/.06/.06					
		GR	44/44/44	118.21***/108.10***/124.05***	.90/.91/.90	.98/.99/.98	.91/.94/.91	.05/.04/.05					
		GA	252/252/252	507.67***/589.09***/610.95***	.90/.90/.90	.97/.95/.95	.73/.71/.66	.04/.05/.05					
2	Konfigurale Invarianz	DB	35/35/35	153.12***/173.43***/209.53***	.91/.90/.90	.98/.98/.98	.87/.87/.81	.06/.05/.05					
		DK	54/54/54	241.86***/269.77***/319.88***	.90/.90/.90	.98/.98/.97	.80/.80/.73	.05/.05/.06					
		WV	42	112.24***	.92	.97	.97	.05					
		AS											
		GP	162	364.14***	.91	.96	.93	.04					
3	Metrische Invarianz	GR	132	442.14***	.92	.96	.89	.04					
		GA	756	707.71***	.91	.95	1.01	.03					
		DB	105	353.59***	.91	.95	.91	.05					
		DK	162	230.24***	.90	.96	.97	.04					
		WV	54	152.41***	.91	.96	.96	.05					
		AS	0										
		GP	184	473.69***	.90	.95	.90	.04					
		GR	152	574.62***	.91	.96	.85	.04					
		GA	802	920.50***	.90	.94	.96	.03					
		DB	123	460.77***	.91	.96	.88	.04					
3 vs. 2		DK	184	334.37***	.90	.95	.94	.04					
		WV							12	40.17*	.01	.01	.01
		AS							22	109.55***	.01	.01	.03
		GP							20	132.48***	.01	<.001	.04
		GR							46	212.79***	.01	.01	.05
		GA							18	107.18***	<.001	<.001	.03
		DB							22	104.13***	<.001	.01	.03
		DK											

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	ΔGH	ANCI
4a	Skalare Invarianz	WV	66	164,14***	.90	.95	.96	.05					
		AS	3	17,08	1,00	.99	.99	.05					
		GP	206	476,45***	.90	.95	.90	.04					
		GR	172	686,29***	.90	.95	.82	.04					
		GA	848	1 015,82***	.91	.90	.94	.04					
		DB	141	693,40***	.89	.88	.81	.07					
		DK	206	510,21***	.90	.95	.89	.04					
		WV							12	11,73	.01	.01	<.001
		AS							3				
		GP							22	2,76	<.001	<.001	<.001
4b	Partiell skalare Invarianz	GR							20	111,67***	.01	.01	.03
		GA							46	95,32***	<.001	.04	.02
		DB ^b							18	232,63***	.02	.08	.07
		DK							22	175,84***	<.001	<.001	.05
		WV	139	634,17***	.91	.96	.83	.04					
		AS											
		GP											
		GR											
		GA											
		DB											
5a	Invarianz der Fehlervarianzen	WV	80	553,73***	.89	.90	.84	.07					
		AS	7	21,18	1,00	.98	.99	.06					
		GP	230	952,19***	.90	.88	.76	.06					
		GR	194	768,83***	.90	.95	.81	.04					
		GA	896	2 199,75***	.90	.97	.61	.02					
		DB											
		DK	230	892,15***	.85	.80	.78	.08					
		WV							14	389,59***	.01	.05	.12
		AS							4	4,1	<.001	.01	<.001
		GP							24	475,74***	<.001	.07	.14
5a vs. 4a		GR							22	82,54*	<.001	<.001	.01
		GA							48	1 183,93***	.01	<.001	.33
		DB							24	381,94***	.05	.15	.11
		DK											

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Adf	$\Delta\chi^2$	ACFI	AGH	ANCI
5b	Partielle Invarianz der Fehlervarianzen	WV ^c	78	428.07***	.90	.95	.88	.05					
		AS											
		Gp ^d	226	501.86***	.90	.91	.90	.05					
		GR											
5b vs. 4a		GA											
		DB	157	701.63***	.90	.95	.81	.04					
		DK ^e	226	554.60***	.90	.97	.88	.03	12	263.93***	<.001	<.001	.08
		WV											
5b vs. 4b		AS							20	25.41	<.001	.04	<.001
		GP											
		GR											
		GA											
		DB							20	44.39*	<.001	<.001	.01
		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
		DB							18	67.46*	.01	.01	.02

Anmerkungen. $N_{IDS} = 1\,266-1\,330$; WV = Wahrnehmung Visuell; AS = Aufmerksamkeit Selektiv; GP = Gedächtnis Phonologisch; GR = Gedächtnis Räumlich-Visuell; GA = Gedächtnis Auditiv; DB = Denken Bildlich; DK = Denken Konzeptuell; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = McDonalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^aModell 1: Einzelgruppenanalysen für Altersgruppen; Modelle 2–5: Multigruppenvergleiche zwischen Altersgruppen. Bei AS entspricht Modell 1 metrischer Invarianz.

^bFreie Schätzung des y-Achsenabschnitts von Item 6 sowie nachfolgend freie Schätzung der Fehlervarianz von Item 6.

^cFreie Schätzung der Fehlervarianz von Item 1.

^dFreie Schätzung der Fehlervarianzen von Item 1 und 6.

^eFreie Schätzung der Fehlervarianzen von Item 1 und 12.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Tabelle 3. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Alter in der allgemeinen Intelligenz in den IDS-P und IDS

Stich- probe	Mo- dell ^a	Modellspezifikation	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
IDS-P	1	Ausgangsmodell	14/14/14	35.98**/31.88**/12.33	.90/.90/1.00	.99/.98/1.00	.95/.97/1.00	.06/.07/.01					
	2	3;0-3;11/4;0-4;11/5;0-5;11											
	3	Konfigurale Invarianz	42	80.19**	.90	.98	.97	.04					
	4	Metrische Invarianz	54	93.27**	.90	.99	.97	.03	12	13.08*	<.001	<.001	<.001
	5	Skalare Invarianz	66	94.00**	.92	.98	.98	.03	12	0.73	<.001	.01	<.001
	6	Invarianz der Fehler- varianzen	80	98.07***	.95	.99	.99	.02	14	4.07	<.001	<.001	<.001
IDS	1	Ausgangsmodell	14/14/14	35.16**/18.89/16.03	.92/.98/.99	.99/1.00/1.00	.97/1.00/1.00	.06/.03/.02					
	2	5;0-6;11/7;0-8;11/9;0-10;11											
	3	Konfigurale Invarianz	42	70.08**	.95	1.00	.99	.02					
	4	Metrische Invarianz	54	79.89**	.96	.99	.99	.02	12	9.81	<.001	.01	<.001
	5	Skalare Invarianz	66	82.30**	.97	1.00	.99	.01	12	2.41	<.001	<.001	<.001
	6	Invarianz der Fehler- varianzen	80	95.76***	.97	1.00	.99	.01	14	13.46*	<.001	<.001	<.001

Anmerkungen. $N_{IDS-P} = 700$; $N_{IDS} = 1\,330$; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = Mc Donalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^a Modell 1: Einzelgruppenanalysen für Altersgruppen; Modelle 2–5: Multigruppenvergleiche zwischen Altersgruppen.

* $p < .05$; ** $p < .01$; *** $p < .001$.

ESM 2: Messinvarianzanalysen über das Geschlecht

Tabelle 1. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Geschlecht in den Untertests der IDS-P

Modell ^a	spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ΔCFI	ΔGH	ΔNCI
1	Ausgangsmodell Jungen/Mädchen	WV	20/20	91.73***/77.74***	.91/.93	.98/.98	.95/.96	.06/.06					
		AS	0/0										
		GP	90/90	334.81***/406.86***	.90/.90	.96/.96	.83/.79	.06/.06					
		GR	35/35	99.26***/78.81***	.92/.95	.98/.98	.95/.97	.06/.06					
		GA	35/35	60.75***/102.77***	.96/.92	.98/.98	.98/.95	.05/.06					
		DB	54/54	190.18***/131.19***	.90/.91	.97/.97	.90/.94	.06/.06					
2	Konfigurale Invarianz	DK	44/44	62.95***/77.58***	.98/.96	.99/.98	.99/.98	.04/.05					
		WV	40	169.47***	.92	.95	.91	.07					
		AS											
		GP	180	767.30***	.90	.96	.66	.04					
		GR	70	197.81***	.93	.97	.91	.05					
		GA	70	163.53***	.94	.98	.94	.04					
3a	Metrische Invarianz	DB	108	350.60***	.90	.96	.84	.05					
		DK	88	166.69***	.95	.98	.95	.04					
		WV	47	174.16***	.92	.96	.91	.06					
		AS	0										
		GP	194	785.88***	.80	.89	.65	.07					
		GR	79	201.41***	.93	.96	.92	.05					
	3a vs. 2	GA	79	169.50***	.94	.98	.94	.04					
		DB	119	359.50***	.90	.95	.84	.05					
		DK	98	169.44***	.95	.98	.95	.03	7	4.69	<.001	<.001	<.001
		WV											
		AS											
		GP							14	18.58*	.10	.07	.01
		GR							9	3.60	<.001	.01	<.001
		GA							9	5.97	<.001	<.001	<.001
		DB							11	8.90	<.001	.01	<.001
		DK							10	2.75	<.001	<.001	<.001

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
3b	Partiell metrische Invarianz	WV											
		AS											
		GP ^b	193	781.44***	.90	.96	.66	.04					
		GR											
		GA											
		DB											
		DK											
		WV											
		AS							13	14.14	<.001	<.001	<.001
		GP											
4a	Skalare Invarianz	GR											
		GA											
		DB											
		DK											
		WV	54	184.14***	.92	.95	.91	.06					
		AS	2	15.20	1.00	.99	.99	.06					
		GP											
		GR	88	289.67***	.94	.97	.87	.04					
		GA	88	181.50***	.94	.97	.94	.04					
		DB	130	360.49***	.90	.95	.85	.05					
4a vs. 3a		DK	108	256.89***	.95	.98	.90	.03					
		WV							7	9.98	<.001	.01	<.001
		AS							2				
		GP											
		GR							9	88.26*	<.001	<.001	.05
		GA							9	12.00	<.001	.01	<.001
		DB							11	0.99	<.001	<.001	<.001
		DK							10	87.45*	<.001	<.001	.05
4b	Partiell skalare Invarianz	WV											
		AS											
		GP	206	866.91***	.90	.96	.62	.04					
		GR											
		GA											
		DB											
		DK											
		WV											
		AS											
		GP							13	85.47*	<.001	<.001	.04
4b vs. 3b		GR											
		GA											
		DB											
		DK											

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
5a	Invarianz der Fehlervarianzen	WV	62	192.80***	.92	.95	.91	.06					
		AS	4	18.77	1.00	.99	.99	.06					
		GP											
		GR	98	335.67***	.93	.97	.84	.04					
		GA	98	194.48***	.93	.97	.93	.04					
		DB	142	417.50***	.90	.96	.82	.04					
		DK	119	309.96***	.94	.98	.87	.03					
		WV							8	8.66	<.001	<.001	<.001
		AS							2	3.57	<.001	<.001	<.001
		GP							10	46.00*	.01	<.001	.03
5b	Partielle Invarianz der Fehlervarian- zen	GR							10	12.98	.01	<.001	.01
		GA							12	57.01*	<.001	<.001	.03
		DB							11	53.07*	.01	<.001	.03
		DK											
		WV											
		AS											
		GP											
		GR	220	909.77***	.90	.96	.61	.04					
		GA											
		DB											
5b vs. 4b		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
		DB											
		DK											
		WV											
		AS											
5b vs. 4b		GP											
		GR											
		GA											
		DB											
		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
5b vs. 4b		DB											
		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
		DB											
		DK											
		WV											
5b vs. 4b		AS											
		GP											
		GR											
		GA											
		DB											
		DK											
		WV											
		AS											
		GP											
		GR											
5b vs. 4b		GA											
		DB											
		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
		DB											
		DK											

Anmerkungen. $N_{IDS,F} = 627-700$; WV = Wahrnehmung Visuell; AS = Aufmerksamkeit Selektiv; GP = Gedächtnis Phonologisch; GR = Gedächtnis Räumlich-Visuell; GA = Gedächtnis Auditiv; DB = Denken Bildlich; DK = Denken Konzeptuell; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = McDonalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^aModell 1: Einzelgruppenanalysen für Jungen und Mädchen; Modelle 2-5: Multigruppenvergleiche zwischen Jungen und Mädchen. Bei AS entspricht Modell 1 metrischer Invarianz.

^bFreie Schätzung der Faktorladung von Item 7 sowie nachfolgend freie Schätzung des y-Achsenabschnitts und der Fehlervarianz von Item 7.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Tabelle 2. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Geschlecht in den Untertests der IDS

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
1	Ausgangsmodell Jungen/Mädchen	WV	14/14	52.69***/57.20***	.90/.86	.98/.99	.97/.97	.07/.06					
		AS	0/0										
		GP	54/54	215.76***/198.94***	.90/.90	.97/.97	.88/.90	.06/.06					
		GR	44/44	194.58***/169.27***	.90/.90	.97/.97	.89/.91	.06/.06					
		GA	252/252	740.24***/796.14***	.90/.90	.95/.97	.69/.67	.05/.04					
		DB	35/35	128.27***/116.37***	.90/.90	.98/.98	.93/.94	.06/.06					
		DK	54/54	162.14***/142.36***	.89/.90	.97/.97	.92/.94	.06/.06					
2	Konfigurale Invarianz	WV	28	239.89***	.90	.97	.92	.06					
		AS											
		GP	108	731.24***	.90	.96	.79	.05					
		GR	88	502.09***	.90	.95	.86	.06					
		GA	504	1 536.38***	.90	.96	.68	.03					
		DB	70	301.04***	.85	.97	.92	.05					
		DK	108	524.81***	.90	.96	.85	.05					
3a	Metrische Invarianz	WV	34	250.35***	.86	.97	.92	.06					
		AS	0										
		GP	119	830.07***	.79	.91	.77	.07					
		GR	98	641.36***	.75	.92	.82	.07					
		GA	527	1 556.06***	.90	.96	.68	.03					
		DB	79	337.14***	.85	.96	.91	.05					
		DK	119	790.47***	.82	.93	.78	.06					
		WV							6	10.46*	.04	<.001	<.001
		AS							11	98.83**	.11	.05	.02
		GP							10	139.27**	.15	.03	.04
		GR							23	19.68	<.001	<.001	<.001
		GA							9	36.10	<.001	.01	.01
3b	Partiell metrische Invarianz	DK							11	265.66***	.08	.03	.07
		WV ^b	33	249.22***	.90	.97	.92	.06					
		AS											
		GP ^c	117	929.54***	.90	.95	.74	.05					
		GR ^d	97	534.53***	.90	.96	.85	.05					
		GA											
		DB											
		DK ^e	117	573.47***	.90	.95	.84	.05					
		WV							5	9.33	<.001	<.001	<.001
		AS							9	198.30***	<.001	.01	.05
		GP							9	32.44	<.001	<.001	.01
3b vs. 2		GR											
		GA											
		DB							9	48.66*	<.001	.01	.01

Modell ^a	Modell- spezifikation	Unter- test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ΔNCI
4a	Skalare Invarianz	WV	2	15.46	.99	.99	.99	.06					
		AS											
		GP											
		GR											
		GA	550	1 619.48***	.90	.96	.67	.03					
		DB	88	357.20***	.84	.96	.90	.05					
		DK											
		WV							2				
		AS											
		GP											
4b	Partiell skalare Invarianz	GR								63.42*	<.001	<.001	.01
		GA							9	20.06	.01	<.001	.01
		DB											
		DK											
		WV	38	253.73***	.90	.96	.92	.06					
		AS											
		GP	126	987.67***	.90	.95	.72	.05					
		GR	106	546.12***	.90	.95	.85	.05					
		GA											
		DB											
4b vs. 3b		DK	126	631.26***	.90	.95	.83	.05	5	4.51	<.001	.01	<.001
		WV											
		AS							9	58.13*	<.001	<.001	.02
		GP							9	11.59	<.001	.01	<.001
		GR											
		GA											
		DB											
		DK							9	57.79*	<.001	<.001	.01
		WV											
		AS											
5a	Invarianz der Fehlervarianzen	WV	4	17.52	.98	.99	.99	.06					
		AS											
		GP											
		GR											
		GA	574	1 660.19***	.90	.96	.66	.03					
		DB	98	552.69***	.83	.91	.84	.07					
		DK											
		WV							2	2.06	.01	<.001	<.001
		AS											
		GP											
5a vs. 4a		GR							24	40.71*	<.001	<.001	.01
		GA							10	195.49***	.01	.05	.06
		DB											
		DK											
		WV											
		AS											
		GP											
		GR											
		GA											
		DB											

Modell ^a	Modell-spezifikation	Unter-test	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
5b	Partielle Invarianz der Fehlervarianzen	WV	44	407.97***	.90	.96	.87	.06					
		AS											
		GP	136	1 087.55***	.90	.95	.70	.05					
		GR	116	985.05***	.90	.95	.72	.05					
		GA											
5b vs. 4a		DB ^f	97	389.05***	.83	.95	.90	.05					
		DK	136	665.12***	.90	.97	.82	.04					
		WV											
		AS											
		GP											
5b vs. 4b		GR											
		GA											
		DB											
		DK							9	31.85*	.01	.01	<.001
		WV							6	154.24**	<.001	<.001	.05
		AS							10	99.88**	<.001	<.001	.02
		GP							10	438.93***	<.001	<.001	.13
		GR											
		GA											
		DB							10	33.86*	<.001	<.001	.01

Anmerkungen. $N_{IDS} = 1\,266-1\,330$; WV = Wahrnehmung Visuell; AS = Aufmerksamkeit Selektiv; GP = Gedächtnis Phonologisch; GR = Gedächtnis Räumlich-Visuell; GA = Gedächtnis Auditiv; DB = Denken Bildlich; DK = Denken Konzeptuell; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = Mc Donalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^aModell 1: Einzelgruppenanalysen für Jungen und Mädchen; Modelle 2-5: Multigruppenvergleiche zwischen Jungen und Mädchen. Bei AS entspricht Modell 1 metrischer Invarianz.

^bFreie Schätzung der Faktorladung von Item 2 sowie nachfolgend freie Schätzung des y-Achsenabschnitts und der Fehlervarianz von Item 2.

^cFreie Schätzung der Faktorladungen von Item 7 und 8 sowie nachfolgend freie Schätzung der y-Achsenabschnitte und der Fehlervarianzen von Item 7 und 8.

^dFreie Schätzung der Faktorladung von Item 2 sowie nachfolgend freie Schätzung des y-Achsenabschnitts und der Fehlervarianz von Item 2.

^eFreie Schätzung der Faktorladungen von Item 3 und 10 sowie nachfolgend freie Schätzung der y-Achsenabschnitte und der Fehlervarianzen von Item 3 und 10.

^fFreie Schätzung der Fehlervarianz von Item 10.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Tabelle 3. Einzelgruppenanalysen und Multigruppenvergleiche: Messinvarianz über das Geschlecht in der allgemeinen Intelligenz in den IDS-P und IDS

Stich- probe	Modell ^a	Modellspezifikation	df	χ^2	CFI	GH	NCI	RMSEA	Δdf	$\Delta\chi^2$	ACFI	AGH	ANCI
IDS-P	1	Ausgangsmodell Jungen/Mädchen	14/14	24.70*23.58	.94/.95	.99/.99	.98/.99	.05/.04					
2		Konfigurale Invarianz	28	48.28*	.94	.99	.99	.03					
3		Metrische Invarianz	34	54.11*	.94	.99	.99	.03					
4a		3 vs. 2 Skalare Invarianz	40	98.77***	.84	.97	.96	.05	6	5.83	<.001	<.001	<.001
4b		4a vs. 3 Partiell skalare Invarianz ^b	38	62.65**	.93	.99	.98	.03	6	44.66***	.10	.02	.03
5		4b vs. 3 Partielle Invarianz der Fehlervarianzen	43	68.07**	.93	.99	.98	.03	4	8.54	.01	<.001	.01
		5 vs. 4b							5	5.42	<.001	<.001	<.001
IDS	1	Ausgangsmodell Jungen/Mädchen	14/14	12.87/36.54**	1.00/.93	1.00/.99	1.00/.98	.00/.05					
2		Konfigurale Invarianz	28	49.42**	.97	1.00	.99	.02					
3a		Metrische Invarianz	34	64.60**	.95	.99	.99	.03					
3b		3a vs. 2 Partiell metrische Invarianz ^c	33	55.58**	.97	.99	.99	.03	6	15.18*	.02	.01	<.001
4		3b vs. 2 Partiell skalare Invarianz	38	77.34**	.96	.98	.99	.04	5	6.16*	<.001	.01	<.001
5		4 vs. 3b Partielle Invarianz der Fehlervarianzen	44	82.21**	.95	.98	.99	.04	5	21.76*	.01	.01	<.001
		5 vs. 4							6	4.87*	.01	<.001	<.001

Anmerkungen. $N_{IDS-P} = 700$; $N_{IDS} = 1\,330$; CFI = Comparative Fit Index; GH = Gamma Hat; NCI = Mc Donalds Non-Centrality Index; RMSEA = Root Mean Square Error of Approximation.

^aModell 1: Einzelgruppenanalysen für Jungen und Mädchen; Modelle 2-5: Multigruppenvergleiche zwischen Jungen und Mädchen.

^bFreie Schätzung der y-Achsenabschnitte der Untertests *Aufmerksamkeit Selektiv* (AS) und *Denken Bildlich* (DB) und nachfolgend freie Schätzung der Fehlervarianzen dieser beiden Untertests.

^cFreie Schätzung der Faktorladung des Untertests *Denken Konzeptuell* (DK) und nachfolgend freie Schätzung des y-Achsenabschnitts und der Fehlervarianz dieses Untertests.

* $p < .05$, ** $p < .01$; *** $p < .001$.

APPENDIX B: Studie 2

Schweizer, F., Grieder, S., Büniger, A. & Grob, A. Vergleich von Intelligenztestleistungen bei Monolingualen, Bilingualen und Fremdsprachigen in den Intelligence and Development Scales–2 (IDS-2). *Manuskript zur Publikation eingereicht bei Diagnostica.*

Stand: 13.08.2018

IDS-2 bei Monolingualen, Bilingualen und Fremdsprachigen

Originalarbeit

Vergleich von Intelligenztestleistungen bei Monolingualen, Bilingualen und Fremdsprachigen in den Intelligence and Development Scales–2 (IDS-2)

Florine Schweizer, Silvia Grieder, Anette Büniger und
Alexander Grob

Universität Basel, Fakultät für Psychologie,
Missionsstrasse 62, 4055 Basel, Schweiz
Telefon: +41 (0)61 207 06 37
Email: florine.schweizer@unibas.ch

Zusammenfassung: Die Studie untersucht Mittelwertdifferenzen bei Monolingualen, Bilingualen und Fremdsprachigen in der allgemeinen Intelligenz sowie 7 Faktoren mit je 2 Untertests mit den *Intelligence and Development Scales-2* (IDS-2). Die Stichprobe besteht aus je 143 miteinander gepaarten monolingualen, bilingualen und fremdsprachigen Kindern und Jugendlichen der Normierungs- und Validierungsstichprobe ($M_{Alter} = 12.33$, $SD_{Alter} = 4.31$; 49 % Jungen). Mit Ausnahme der allgemeinen Intelligenz (7–10 %) und des Faktors *Denken Verbal* (15 %) wird in den Faktoren maximal 4 % der Varianz durch die Gruppenzugehörigkeit aufgedeckt: Die Fremdsprachigen erzielten in der allgemeinen Intelligenz, im Kurzzeit- und Langzeitgedächtnis sowie im abstrakten und verbalen Denken tiefere Werte als die Monolingualen, während die Unterschiede zwischen Bilingualen und Monolingualen nicht bedeutsam waren. Die Befunde implizieren, dass ein möglicher sprachlicher Nachteil für Personen mit fremdsprachigem Hintergrund bei der Interpretation der Intelligenz-Testergebnisse in den IDS-2 berücksichtigt werden muss.

Schlüsselwörter: Intelligenz; IDS-2; monolingual; bilingual; fremdsprachig

Intelligence comparison between monolingual, bilingual, and non-native speaking children and adolescents in the Intelligence and Development Scales-2 (IDS-2)

Abstract: This study investigates mean level differences in monolinguals, bilinguals, and non-native speakers in general intelligence and 7 factors with 2 subtests each with the *Intelligence and Development Scales-2* (IDS-2). The sample consists of 143 monolingual, bilingual, and non-native speaking children and adolescents of the standardization and validation study, who were matched with each other ($M_{age} = 12.33$, $SD_{age} = 4.31$; 49 % males). With the exception of general intelligence (7–10 %) and the factor *Verbal Reasoning* (15 %), a maximum of 4 % in the factors is explained by group: Compared to monolinguals, non-native speakers scored lower on general intelligence, short-term and long-term memory, as well as abstract and verbal reasoning, whereas difference scores between monolinguals and bilinguals were negligible. Results indicate that possible linguistic disadvantages for people with a non-native background have to be considered when interpreting IDS-2 intelligence test results.

Keywords: intelligence; IDS-2; monolingual; bilingual; non-native background

Einleitung

In der psychologischen Praxis gehören Intelligenztests zu den am häufigsten eingesetzten Testverfahren (Goldstein, Princiotta & Naglieri, 2015). Intelligenztestleistungen sind prädiktiv für verschiedene Lebensaspekte wie beispielsweise akademische Leistungen, sozio-ökonomischer Status, Gesundheit und Langlebigkeit (z.B. Deary, 2009; Roth et al., 2015; Strenze, 2007). Dies mag einer der wichtigsten Gründe dafür sein, dass für nachhaltige Entscheidungen wie beispielsweise Schulungsform oder Fördermassnahmen Intelligenztests verwendet werden. Eine zentrale Frage ist, welche Tests bei fremdsprachigen Personen, die eine andere Muttersprache als die Testsprache sprechen, valide eingesetzt werden können. Fremdsprachige Personen wachsen meist mehrsprachig auf, was vielfach mit einer altersverzögerten Sprachentwicklung sowie einem geringer ausgeprägten Sprachverständnis und Wortschatz in der Landessprache einhergeht (Reich et al., 2002).

Für fremdsprachige Personen bestehen bei standardisierten Intelligenzverfahren besondere Herausforderungen (Weiss et al., 2006). Testverfahren mit hohen sprachlichen Anforderungen können zu Problemen beim Instruktionsverständnis und bei der Produktion einer (verbalen) Antwort führen (Daseking & Petermann, 2015; Weiss et al., 2006). Es ist somit wahrscheinlich, dass fremdsprachige Personen nicht ihr ganzes Potential zeigen können und in ihren Fähigkeiten unterschätzt werden (Hagmann-von Arx, Petermann & Grob, 2013). Allgemein ist davon auszugehen, dass ihre Testwerte geringer ausfallen, je sprachgebundener die Aufgaben eines Verfahrens sind (Flanagan & Ortiz, 2001). Bei der Interpretation ist es folglich wichtig, die Testergebnisse mit den sprachlichen Anforderungen eines Tests in Beziehung zu setzen (Schölmerich, Leyendecker, Citlak, Caspar & Jäkel, 2008). Diese variieren stark zwischen Testverfahren (Hagmann-von Arx, Meyer & Grob, 2008).

Die vorliegende Studie untersucht bei den *Intelligence and Development Scales-2* (IDS-2; Grob & Hagmann-von Arx, 2018), ob sich Mittelwertdifferenzen zwischen deutsch- und fremdsprachigen Personen in der Intelligenz zeigen. Die IDS-2 basieren auf den *Intelligence and Development Scales* für Kinder von 5;0 bis 10;11 Jahren (IDS; Grob, Meyer & Hagmann-von Arx, 2013); letztere wurden zusätzlich als *Intelligence and Development Scales-Preschool* (IDS-P; Grob, Reimann, Gut & Frischknecht, 2013) für Kinder von 3;0 bis 5;11 Jahren weiterentwickelt. Die IDS-2 erfassen kognitive Funktionen (*Intelligenz, Exekutive Funktionen*) und allgemeine Entwicklungsfunktionen (*Psychomotorik, Sozial-Emotionale Kompetenz, Schulische Kompetenzen, Arbeitshaltung*) bei Kindern und Jugendliche zwischen 5;0 und 20;11 Jahren. Die Intelligenz wird in Anlehnung an die Cattell-Horn-Carroll-Theorie (CHC-Theorie; McGrew, 2005) auf drei Ebenen erfasst: Die allgemeine Intelligenz besteht aus sieben Faktoren, welche aus jeweils zwei Untertests zusammengesetzt sind. Während die ersten beiden Faktoren basale Wahrnehmungs- und Verarbeitungsaspekte erfassen, nehmen die nachfolgenden Faktoren hinsichtlich Komplexität als auch—zumindest teilweise—Sprachgebundenheit zu: Der Faktor *Verarbeitung Visuell* erfasst visuelle Wahrnehmungs- und Diskriminationsfähigkeiten. *Verarbeitungsgeschwindigkeit* erfasst die Fähigkeit zur automatischen und flüssigen Ausführung von kognitiven Aufgaben sowie selektive Aufmerksamkeit. *Kurzzeitgedächtnis Auditiv* und *Kurzzeitgedächtnis Räumlich-Visuell* erheben die Fähigkeit zur Speicherung und zum Abruf von verbalen respektive visuell-räumlichen Informationen im und aus dem Kurzzeit- und Arbeitsgedächtnis; *Langzeitgedächtnis* misst die Fähigkeit zur Speicherung und zum Abruf von verbalen und visuell-räumlichen Informationen im und aus dem Langzeitgedächtnis. *Denken Abstrakt* erfasst fluides logisches und schlussfolgerndes Denken und *Denken Verbal* die kristalline verbale Wissensaneignung und -anwendung.

Im Rahmen der Normierungs- und Validierungsstudien der IDS-2 wurden verschiedene Untersuchungen zur Konstrukt-, Kriteriums- und differentiellen Validität durchgeführt (vgl. Grob & Hagmann-von Arx, 2018). Bislang liegen allerdings keine Befunde zu Mittelwertvergleichen zwischen verschiedenen Sprachgruppen vor. Es ist relevant, diese Forschungslücke im Sinne der differentiellen Validität zu füllen. Die Validierung von Intelligenztests ist essenziell, um deren Gültigkeit und Generalisierbarkeit zu gewährleisten (Schmidt-Atzert & Amelang, 2012). Dies gilt insbesondere für Neuentwicklungen und -normierungen (Daseking, Lipsius, Petermann & Waldmann, 2008).

Forschungsbefunde

Die Vorgängerversionen der IDS-2, die IDS-P und IDS, zeigen lediglich minime Differenzen zwischen deutsch- und fremdsprachigen Kindern: In der allgemeinen Intelligenz, welche mehrheitlich fluide Intelligenzaspekte erfasst, und in den Untertests zu visueller Wahrnehmung, selektiver Aufmerksamkeit, phonologischem und visuell-räumlichem Kurzzeitgedächtnis sowie bildlichem und konzeptuellem Denken zeigen sich keine Mittelwertdifferenzen. Es finden sich jedoch Unterschiede zuungunsten von Fremdsprachigen im Untertest *Gedächtnis Auditiv*, welcher Aspekte des verbalen Langzeitgedächtnisses erfasst; für die IDS mit einem kleinen ($d = 0.45$), für die IDS-P mit einem großen ($d = 0.91$) Effekt (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013).

Weitere deutschsprachige Testverfahren zeigen ebenfalls Differenzen zwischen deutsch- und fremdsprachigen Kindern und Jugendlichen, wobei diese je nach Zusammensetzung und theoretischem Hintergrund des Tests unterschiedlich ausfallen: In der allgemeinen Intelligenz zeigt die *Wechsler Intelligence Scale for Children–Vierte Auflage* (WISC-IV; Petermann & Petermann, 2011), welche sowohl fluide als auch kristalline Intelligenzaspekte erfasst, tiefere Werte für Fremdsprachige mit kleinen Effekten ($d = 0.27$ – 0.39 ; Daseking et al., 2008¹; Hagmann-von Arx et al., 2013). Während sich ebenfalls in den Normierungsstichproben der *Snijders-Oomen nonverbalen Intelligenztests* (SON-R 2½-7 und SON-R 6-40; Tellegen, Laros & Petermann, 2007, 2012) kleine bis mittlere Unterschiede zuungunsten von Fremdsprachigen finden ($d = 0.46$ – 0.54 ; Tellegen et al., 2007, 2012), sind solche in weiteren Studien nicht vorhanden (Janke, Daseking & Petermann, 2008; Toussaint, Heinze, Lipsius & Petermann, 2012).

Auf Faktoren- und Untertestebene zeigen sich keine Unterschiede zwischen Deutsch- und Fremdsprachigen in den Indizes *Verarbeitungsgeschwindigkeit* und *Arbeitsgedächtnis* der WISC-IV (Daseking et al., 2008; Hagmann-von Arx et al., 2013). Weiter finden sich keine Unterschiede in der *Handlungs-* und *Denkskala* des SON-R 2½-7 (Janke et al., 2008) sowie in den Untertests des SON-R 6-40 (Toussaint et al., 2012). Fremdsprachige erreichen jedoch tiefere Werte in der WISC-IV im Index *Wahrnehmungsgebundenes Logisches Denken* ($d = 0.34$) und den Untertests *Mosaik-Test* ($d = 0.38$) und *Bildkonzepte* ($d = 0.27$)—nicht jedoch *Matrizen-Test* ($d = 0.16$)—sowie im Index *Sprachverständnis* ($d = 0.47$) und den Untertests *Gemeinsamkeiten finden* ($d = 0.39$), *Wortschatz-Test* ($d = 0.43$) und *Allgemeines Verständnis* ($d = 0.44$; Daseking et al., 2008).

Insgesamt weisen Forschungsbefunde zu anderen Testverfahren darauf hin, dass sich Unterschiede zwischen deutsch- und fremdsprachigen Personen je nach Zusammensetzung des Tests zeigen. Die Effektstärken bewegen sich dabei mehrheitlich im kleinen Bereich.

¹ Bei Stichprobenpaarung nach Alter, Geschlecht und sozio-ökonomischem Status

Forschungsziele und Hypothesen

Das Ziel der vorliegenden Studie ist es, Mittelwertdifferenzen in der Intelligenz zwischen verschiedenen Familiensprachgruppen in den IDS-2 zu untersuchen, um deren Validierung im Rahmen der differentiellen Validität zu erweitern. Der Großteil der bisherigen Studien hat jeweils Unterschiede zwischen deutsch- und fremdsprachigen Kindern und Jugendlichen untersucht, wobei nicht deutlich hervorgeht, ob die Unterschiede sowohl für Fremd- als auch für Zweisprachige gefunden werden können. Daher besteht ein weiteres Ziel dieser Studie darin, innerhalb der fremdsprachigen Personen zwei Substichproben zu unterscheiden: Zum einen *Fremdsprachige*, welche eine andere Muttersprache als Deutsch sprechen; zum anderen *Bilinguale (Zweisprachige)*, welche sowohl Deutsch als auch eine andere Sprache sprechen. Als Vergleichsgruppe dienen *Monolinguale*, welche Deutsch als einzige Muttersprache haben.

In Anlehnung an die oben dargestellten Befunde erwarten wir Mittelwertunterschiede zuungunsten von fremdsprachigen im Vergleich zu monolingualen Kindern und Jugendlichen in den sprachgebundenen Intelligenzskalen der IDS-2: Zum einen erwarten wir tiefere Mittelwerte für Fremdsprachige in der allgemeinen Intelligenz, da sich diese sowohl aus fluiden wie auch kristallinen Aspekten zusammensetzt. Weiter erwarten wir tiefere Werte für Fremdsprachige in den sprachgebundenen Faktoren *Langzeitgedächtnis* und *Denken Verbal* sowie den zugehörigen Untertests *Geschichten nacherzählen* und *Bild beschreiben* respektive *Kategorien nennen* und *Gegenteile nennen*. In den anderen Faktoren und Untertests erwarten wir indes keine Unterschiede zwischen den zwei Gruppen, da diese weniger sprachlastig sind. Zusätzlich untersuchen wir, ob für Bilinguale ebenfalls Unterschiede im Vergleich zu Monolingualen bestehen und ob potenzielle Unterschiede zwischen Monolingualen und Fremdsprachigen grösser sind als solche zwischen Mono- und Bilingualen, da Fremd- im Vergleich zu Zweisprachigen weniger Kontakt zur deutschen Sprache haben.

Methoden

Stichprobe

Die Versuchspersonen stammen aus einer Teilstichprobe der Normierungs- und Validierungsstichprobe der IDS-2 (Grob & Hagmann-von Arx, 2018). Die drei Gruppen (Fremdsprachige, Bilinguale, Monolinguale) umfassen jeweils 143 Kinder und Jugendliche und wurden nach Alter, Geschlecht und höchster mütterlicher Ausbildung (Hochschulabschluss vs. kein Hochschulabschluss) gepaart. Jede der Gruppen schließt 70 Jungen (49 %) und 45 Probanden (32 %), deren Mutter über einen Hochschulabschluss verfügt, ein. Die Probanden stammen aus der Schweiz und aus Deutschland. Die *Fremdsprachigen* sind im Durchschnitt 12.46 Jahre alt ($SD = 4.32$) und sprechen eine andere Muttersprache als Deutsch. Als andere Sprachen wurden Albanisch (13 %), Italienisch (8 %), Russisch (8 %), Türkisch (8 %), Bulgarisch (6 %), Serbisch (6 %) und Tamilisch (6 %) sowie verschiedene weitere Sprachen (jeweils < 5 %) angegeben. Die *Bilingualen* sind im Durchschnitt 12.05 Jahre alt ($SD = 4.32$) und sprechen gleichzeitig die Muttersprache Deutsch sowie eine andere Muttersprache. Als zusätzliche Sprachen wurden Englisch (15 %), Russisch (12 %), Türkisch (11 %), Französisch (10 %), Italienisch (10 %) und Spanisch (8 %) sowie verschiedene weitere Sprachen (jeweils < 5 %) angegeben. Die *Monolingualen* sind im Durchschnitt 12.45 Jahre alt ($SD = 4.30$) und sprechen als ausschließliche Muttersprache Deutsch. Die Rekrutierung erfolgte über Spielgruppen, Kindergärten und Schulen sowie private und institutionelle Kontakte. Die Erhebung fand von 2015 bis 2017 statt.

Verfahren

Die IDS-2 umfassen 14 Untertests zur Einschätzung der allgemeinen Intelligenz, von welchen jeweils 2 einen Faktor höherer Ordnung bilden. Die allgemeine Intelligenz kann als *IQ-Profil* (7 Faktoren mit 14 Untertests), *IQ* (7 Untertests) oder *IQ-Screening* (2 Untertests) erfasst werden. Die Untertests *Figuren nachlegen* (geometrische Figuren mit Rechtecken und Dreiecken nachlegen) und *Plättchen legen* (Plättchen nach einer Vorlage nachlegen) bilden den Faktor *Verarbeitung Visuell*. Die Untertests *Zwei Merkmale durchstreichen* (Papageien mit 2 orangen Merkmalen aus Reihen verschiedener Papageien möglichst rasch richtig durchstreichen) und *Kästchen durchstreichen* (Gruppen mit 3 oder 4 Kästchen aus Reihen verschieden großer Gruppen möglichst rasch richtig durchstreichen) bilden den Faktor *Verarbeitungsgeschwindigkeit*. Die Untertests *Zahlen- und Buchstabenreihen nachsprechen* und *Gemischte Zahlen- und Buchstabenreihen nachsprechen* ([gemischte] Zahlen- und Buchstabenreihen vorwärts und rückwärts nachsprechen) ergeben den Faktor *Kurzzeitgedächtnis Auditiv* und die Untertests *Figuren wiedererkennen* und *Rotierte Figuren wiedererkennen* ([rotierte] Figuren speichern und aus einer Auswahl die Formen und Positionen wiedererkennen) den Faktor *Kurzzeitgedächtnis Räumlich-Visuell*. *Geschichte nacherzählen* (eine Geschichte hören und nach mindestens 20 Minuten frei und gestützt erinnern) und *Bild beschreiben* (ein Bild betrachten und nach mindestens 20 Minuten frei und gestützt beschreiben) werden dem Faktor *Langzeitgedächtnis* zugeordnet. Der Faktor *Denken Abstrakt* besteht aus den Untertests *Matrizen ergänzen* (erkennen, wie sich eine Figur verändert und diese Veränderung auf eine weitere Figur anwenden) und *Unpassende Bilder ausschließen* (aus einer Anzahl von Bildern entscheiden, welches Bild nicht zu den anderen passt). Schließlich enthält der Faktor *Denken Verbal* die beiden Untertests *Kategorien nennen* (für Bilder oder Begriffe die korrekte Kategorie nennen) und *Gegenteile nennen* (Gegenteile zu vorgegebenen Wörtern nennen). Die Untertests *Figuren nachlegen*, *Zwei Merkmale durchstreichen*, *Zahlen- und Buchstabenreihen nachsprechen*, *Figuren wiedererkennen*, *Geschichte nacherzählen*, *Matrizen ergänzen* und *Kategorien nennen* fließen in die Berechnung des *IQ* ein. Das *IQ-Screening* besteht aus den Untertests *Matrizen ergänzen* und *Kategorien nennen*. Die Untertests sind mit Wertpunkten ($M = 10$, $SD = 3$) und die Faktoren und die allgemeine Intelligenz mit IQ-Werten ($M = 100$, $SD = 15$) normiert.

Statistisches Vorgehen

Die Analysen wurden mit dem Statistikprogramm R (R Core Team, 2016) vorgenommen. Das Matching wurde mit dem R-Paket MatchIt (Ho, Imai, King & Stuart, 2011) durchgeführt. Mittelwertdifferenzen in der allgemeinen Intelligenz, den Faktoren und Untertests zwischen den Gruppen wurden basierend auf den Normwerten anhand von einfaktoriellen Varianzanalysen ermittelt. Dabei wurden zuerst Globaleffekte über alle Gruppen und anschließend alle möglichen Gruppenvergleiche mittels post-hoc Kontrasten überprüft. Alle p -Werte wurden zusätzlich mit der Korrektur nach Hommel (1988) adjustiert (p_H), um der Alpha-Fehler-Inflation entgegenzuwirken, welche durch multiples Testen entstehen kann. Für die Globaleffekte wurden zusätzlich Effektstärken η^2 berechnet, wobei $\eta^2 \geq .01$ einem kleinen, $\eta^2 \geq .06$ einem mittleren und $\eta^2 \geq .14$ einem großen Effekt entspricht. Für die Mittelwertdifferenzen wurden zusätzlich Effektstärken d berechnet, wobei $d \geq .20$ einem kleinen, $d \geq .50$ einem mittleren und $d \geq .80$ einem großen Effekt entspricht (Cohen, 1988). Differenzen zwischen Gruppen wurden als bedeutsam interpretiert, wenn sowohl Signifikanz ($p_H < .05$) und mindestens kleine Effekte ($\eta^2 \geq .01$) beim Globaleffekt als auch Signifikanz ($p_H < .05$) und mindestens kleine Effekte ($d \geq .20$) beim post-hoc Kontrast erreicht wurden.

Ergebnisse

Die deskriptive Statistik für monolinguale, bilinguale und fremdsprachige Kinder und Jugendliche ist in Tabelle 1 dargestellt. Die Reliabilitäten für die allgemeine Intelligenz ($\alpha = .95-.99$), die Faktoren ($\alpha = .92-.98$) und die Untertests ($\alpha = .85-.97$) liegen im hohen Bereich.

Tabelle 1 hier einfügen

Mittelwertdifferenzen zwischen monolingualen, bilingualen und fremdsprachigen Kindern und Jugendlichen in der Intelligenz der IDS-2 sind in Tabelle 2 dargestellt. Auf Ebene der Globaleffekte erklärte die Gruppenzugehörigkeit 7 bis 10 % der Varianz in der allgemeinen Intelligenz (*IQ-Profil*, *IQ* und *IQ-Screening*; $p_H < .001$). Auf Ebene der Kontraste zeigten Monolinguale signifikant höhere Mittelwerte als Fremdsprachige mit mittleren Effekten ($d = 0.65-0.77$, $p_H < .001$), und Bilinguale zeigten ebenfalls signifikant höhere Mittelwerte als Fremdsprachige mit kleinen bis mittleren Effekten ($d = 0.35-0.52$, $p_H < .001$). Im Vergleich zu Monolingualen zeigten sich kleine Unterschiede zuungunsten von Bilingualen ($d = 0.27-0.30$), die aber nicht signifikant wurden ($p_H = .47-.64$).

Auf Ebene der Globaleffekte erklärte die Gruppenzugehörigkeit 0 bis 15 % der Varianz in den Faktoren und 0 bis 13 % in den Untertests: Von 7 Faktoren lag bei 3 kein Effekt, bei 3 ein signifikanter kleiner und bei 1 ein signifikanter großer Effekt vor; von 14 Untertests lag bei 7 kein, bei 4 ein signifikanter kleiner und bei 3 ein signifikanter mittlerer Effekt zwischen den Gruppen vor. Der Faktor *Verarbeitung Visuell* und die Untertests *Figuren nachlegen* und *Plättchen legen*, der Faktor *Verarbeitungsgeschwindigkeit* und die Untertests *Zwei Merkmale durchstreichen* und *Kästchen durchstreichen*, der Faktor *Kurzzeitgedächtnis Räumlich-Visuell* und der Untertest *Rotierte Figuren wiederkennen* sowie die Untertests *Bild beschreiben* (Faktor *Langzeitgedächtnis*) und *Unpassende Bilder ausschließen* (Faktor *Denken Abstrakt*) wiesen keine bedeutsamen Unterschiede zwischen den Gruppen auf und die Gruppenzugehörigkeit erklärte lediglich 0 bis 2 % der Varianz ($p_H = .07-.51$).

Signifikante kleine Globaleffekte wurden für den Faktor *Kurzzeitgedächtnis Auditiv* (4 % erklärte Varianz, $p_H < .001$) und die Untertests *Zahlen- und Buchstabenreihen nachsprechen* und *Gemischte Zahlen- und Buchstabenreihen nachsprechen* (je 3 % erklärte Varianz, $p_H = .01$), den Faktor *Denken Abstrakt* (3 % erklärte Varianz, $p_H = .01$) und den Untertest *Matrizen ergänzen* (3 % erklärte Varianz, $p_H = .02$), den Faktor *Langzeitgedächtnis* (4 % erklärte Varianz, $p_H < .001$) sowie den Untertest *Figuren wiederkennen* (3 % erklärte Varianz, $p_H = .02$) des Faktors *Kurzzeitgedächtnis Räumlich-Visuell* ersichtlich. Auf Ebene der Kontraste zeigten sich signifikante kleine Unterschiede in *Kurzzeitgedächtnis Auditiv* ($d = 0.49$, $p_H < .001$), *Zahlen- und Buchstabenreihen nachsprechen* ($d = 0.43$, $p_H = .02$), *Gemischte Zahlen- und Buchstabenreihen nachsprechen* ($d = 0.45$, $p_H = .01$), *Denken Abstrakt* ($d = 0.45$, $p_H = .01$), *Matrizen ergänzen* ($d = 0.42$, $p_H = .02$) und *Figuren wiederkennen* ($d = 0.41$, $p_H = .03$) sowie signifikante mittlere Unterschiede in *Langzeitgedächtnis* ($d = 0.50$, $p_H < .001$) zwischen Monolingualen und Fremdsprachigen zuungunsten letzterer. Ein signifikanter mittlerer Globaleffekt wurde für den Untertest *Geschichte nacherzählen* (7 % erklärte Varianz, $p_H < .001$) des Faktors *Langzeitgedächtnis* ersichtlich, indem sich auf Ebene der Kontraste signifikante mittlere Unterschiede zwischen Monolingualen und Fremdsprachigen zuungunsten letzterer zeigten ($d = 0.63$, $p_H < .001$). Schließlich ergaben sich signifikante große Globaleffekte für den Faktor *Denken Verbal* mit 15 % erklärter Varianz ($p_H < .001$) sowie signifikante mittlere Globaleffekte für die Untertests *Kategorien nennen* und *Gegenteile nennen* mit 12 % respektive 13 % erklärter Varianz ($p_H < .001$). Auf Ebene der Kontraste zeigten sich jeweils signifikante große Effekte zwischen Monolingualen und Fremdsprachigen ($d = 0.86-1.01$, $p_H < .001$) sowie signifikante mittlere Effekte zwischen Zwei- und Fremdsprachigen ($d = 0.55-0.66$, $p_H < .001$), jeweils zuungunsten Fremdsprachiger. Die restlichen Kontraste—

darunter alle Kontraste zwischen Mono- und Bilingualen—zeigten entweder keine ($p_H = .65$, $d = -0.09-0.19$) oder kleine, jedoch nicht signifikante Effekte ($p_H = .09-.65$, $d = 0.20-0.38$).

Tabelle 2 hier einfügen

Die Korrelationen zwischen den einzelnen Variablen können dem Elektronischen Supplement ESM entnommen werden.

Diskussion

Das Ziel der vorliegenden Studie war zu prüfen, ob in den IDS-2 Mittelwertdifferenzen in der allgemeinen Intelligenz sowie Faktoren und Untertests der Intelligenz zwischen monolingualen, bilingualen und fremdsprachigen Kindern und Jugendlichen vorliegen. In der allgemeinen Intelligenz erklärte die Gruppenzugehörigkeit zwischen 7 und 10 % der Varianz. Erwartungsgemäß zeigten sich tiefere Mittelwerte für Fremdsprachige im Vergleich zu Monolingualen, und zwar sowohl im *IQ-Profil* als auch im *IQ* und *IQ-Screening*. Weiter waren die Mittelwerte der Fremdsprachigen ebenfalls tiefer als diejenigen der Bilingualen, wobei der Unterschied zwischen Bi- und Monolingualen klein, jedoch nicht signifikant war. Der Befund zum Unterschied zwischen Fremdsprachigen und Monolingualen in der allgemeinen Intelligenz ist mit Studien zur WISC-IV vergleichbar (Daseking et al., 2008; Hagmann-von Arx et al., 2013), während die IDS-P und IDS sowie Studien zum SON-R 2½-7 und SON-R 6-40 keine bedeutsamen Unterschiede zwischen Deutsch- und Fremdsprachigen finden konnten (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013; Janke et al., 2008; Toussaint et al., 2012). Die allgemeine Intelligenz der IDS-2 erfasst ähnlich wie diejenige der WISC-IV zusätzlich zu fluiden kristallinen Intelligenzaspekten und ist somit sprachgebundener als ihre Vorgängerversionen IDS-P und IDS (vgl. Grob & Hagmann-von Arx, 2018). Folglich ist es plausibel, dass monolingual Deutschsprachige die höchsten Werte erzielen, gefolgt von zweisprachigen und zuletzt fremdsprachigen Personen.

Bezüglich Differenzen in den Faktoren und Untertests erwarteten wir tiefere Mittelwerte für fremdsprachige Kinder und Jugendliche im Faktor *Denken Verbal* sowie den Untertests *Kategorien nennen* und *Gegenteile nennen*. Im Faktor und beiden Untertests zeigten sich tiefere Werte für Fremdsprachige im Vergleich zu Mono- sowie im Vergleich zu Bilingualen; weiter war der Unterschied zwischen Bi- und Monolingualen klein, jedoch nicht signifikant. Der Befund zum Unterschied zwischen Fremdsprachigen und Monolingualen ist konsistent mit Studien zum WISC-IV-Index *Sprachverständnis* und den zugehörigen Untertests (Daseking et al., 2008). Die vorliegenden Ergebnisse sind dadurch erklärbar, dass dieser IDS-2-Faktor und die Untertests kristalline sowie sprachgebundene Intelligenzaspekte erheben, was zu Nachteilen führen könnte, je weniger Kontakt mit der deutschen Sprache besteht und folglich je weniger Deutschkenntnisse vorhanden sind. Die erklärte Varianz durch die Gruppenzugehörigkeit war im Faktor *Denken Verbal* (15 %) und den beiden zugehörigen Untertests (12–13 %) im Vergleich zu den anderen Skalen am größten.

Weiter erwarteten wir tiefere Testwerte für Fremdsprachige im Faktor *Langzeitgedächtnis* sowie den Untertests *Geschichte nacherzählen* und *Bild beschreiben*. Während die Fremdsprachigen im Faktor sowie im Untertest *Geschichte nacherzählen* tiefere Werte als Monolinguale zeigten, fanden sich im Untertest *Bild beschreiben* keine bedeutsamen Unterschiede zwischen den Gruppen. Der Befund zu *Geschichte nacherzählen* ist mit den Befunden zu Deutsch- und Fremdsprachigen im Untertest zum verbalen Langzeitgedächtnis der IDS-P und IDS konform (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). Ein Grund für Unterschiede zuungunsten von Fremdsprachigen in *Geschichte nacherzählen* könnte sein, dass der zu Erinnernde Inhalt sprachlich dargeboten und später sprachlich abge-

fragt wird, was hohe Sprachkompetenzen erfordert. Dahingegen könnte ein Grund für fehlende Unterschiede in *Bild beschreiben* sein, dass der Abruf sprachlich sowie visuell und die Darbietung visuell erfolgt, was insofern weniger Sprachkompetenzen voraussetzt. Die Bilingualen erreichten zwar sowohl im Faktor als auch in beiden Untertests leicht tiefere Werte als Monolinguale, die Unterschiede waren jedoch nicht signifikant.

In den anderen Faktoren und Untertests erwarteten wir keine Unterschiede zwischen den Gruppen. Dies konnte für *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* sowie deren Untertests bestätigt werden. Diese Ergebnisse gehen mit Studien einher, welche keine tieferen Werte für fremdsprachige Personen in visuellen Verarbeitungsaspekten und in der Verarbeitungsgeschwindigkeit (Daseking et al., 2008; Hagmann-von Arx et al., 2013; Janke et al., 2008) sowie in den ähnlichen Untertests der IDS-P und IDS (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013) finden konnten. Diese beiden Intelligenzaspekte sind relativ sprachunabhängig (vgl. McGrew, 2005).

Der Faktor *Kurzzeitgedächtnis Räumlich-Visuell* und dessen zweiter Untertest *Rotierte Figuren wiedererkennen* zeigten keine Unterschiede zwischen den Gruppen, wobei sich jedoch tiefere Werte für Fremdsprachige im Vergleich zu Monolingualen im ersten Untertest *Figuren wiedererkennen* zeigten. Dieser Befund ist ähnlich demjenigen der IDS-P und IDS, welche ebenfalls keine Unterschiede im räumlich-visuellen Kurzzeit- respektive Arbeitsgedächtnis finden konnten (Grob, Meyer et al., 2013; Grob, Reimann et al., 2013). Die Untertests (*Rotierte Figuren wiedererkennen*) haben zur Aufgabe, (rotierte) Figuren zu speichern und aus einer Auswahl die Formen und Positionen wiederzuerkennen. Bei beiden Untertests gibt es jeweils drei Aufgabenarten: (rotierte) Figuren wiedererkennen, (rotierte) Figuren unter Distraction der Farbe wiedererkennen und (rotierte) Figuren sowie deren Positionen unter Distraction der Farbe wiedererkennen. Ein Grund für tiefere Werte bei Fremdsprachigen im ersten Untertest könnte sein, dass die Instruktion relativ komplex ist, was möglicherweise zu sprachlichen Verständnisschwierigkeiten führt. Dahingegen könnten die fehlenden Unterschiede im zweiten strukturgleichen Untertest darauf zurückzuführen sein, dass durch eine Wiederholung der Instruktion oder die sehr ähnlichen Aufgaben vom ersten zum zweiten Untertest ein Lerneffekt stattgefunden haben könnte. Dies würde mit dem sogenannten Dynamischen Testen (DT) einhergehen: DT ist ein Ansatz, der vor allem bei Minoritäten eine validere Intelligenzeinschätzung anstrebt, indem bei sogenannten Lerntests Denkhilfen oder Feedbacks in den Testprozess eingebaut werden (Guthke & Wiedl, 1996). Eine Studie mit Kindern mit Migrationshintergrund hat beispielsweise gezeigt, dass diese stärker von einem Lernetest profitieren als muttersprachliche Kinder (Hessels, 1997). Angewendet auf die IDS-2 könnte dies bedeuten, dass der Lerneffekt für Fremdsprachige stärker ausfällt als derjenige für Monolinguale und dies zur Verringerung der Unterschiede im zweiten Untertest führt.

Der Faktor *Kurzzeitgedächtnis Auditiv* (Untertests *Zahlen- und Buchstabenreihen nachsprechen* und *Gemischte Zahlen- und Buchstabenreihen nachsprechen*) zeigte tiefere Werte für Fremdsprachige im Vergleich zu Monolingualen. Dieser Befund stimmt nicht mit Befunden zu den IDS-P und IDS sowie Studien zur WISC-IV überein, welche keine Unterschiede im phonologischen Kurzzeit- respektive Arbeitsgedächtnis finden konnten (Daseking et al., 2008; Grob, Meyer et al., 2013; Grob, Reimann et al., 2013; Hagmann-von Arx et al., 2013). Die IDS-2 beinhalten sowohl Zahlen als auch Buchstaben und sowohl einfache (nur Zahlen oder Buchstaben) wie auch gemischte (sowohl Zahlen als auch Buchstaben) Reihen; weiter müssen die Reihen in einem ersten Teil vorwärts und in einem zweiten Teil rückwärts wiedergegeben werden. Diese Aufgabenstellung erfordert womöglich höhere Sprachkenntnisse als die Vorgängerversionen der IDS-2, was mit Unterschieden zuungunsten von Fremdsprachigen in der vorliegenden Studie einhergeht. Obwohl in der WISC-IV ebenfalls Zahlen vorwärts und rückwärts nachgesprochen (*Zahlen nachsprechen*) sowie Zahlen und Buchstaben geordnet und nachgesprochen (*Buchstaben-Zahlen-Folgen*) werden müssen, lassen sich dort keine Unterschiede finden (Daseking et al., 2008).

Schließlich resultierten ebenfalls tiefere Werte zuungunsten Fremdsprachiger verglichen mit Monolingualen im Faktor *Denken Abstrakt* und im Untertest *Matrizen ergänzen*, nicht jedoch im Untertest *Unpassende Bilder ausschließen*. Während für den WISC-IV-Index *Wahrnehmungsgebundenes Logisches Denken* und zwei der drei Untertests ebenfalls tiefere Werte für Fremd- im Vergleich zu Deutschsprachigen gefunden wurden (Daseking et al., 2008), zeigten die Matrizen-Untertests der WISC-IV und des SON-R 6-40 keine Differenzen (Daseking et al., 2008; Toussaint et al., 2012). Die hier gefundenen Unterschiede entsprechen dem Befund, dass auch bei der Bearbeitung von sprachfreien Intelligenztestaufgaben sprachliche Leistungen erforderlich sind, da die Testsprache sowie der Testprozess verstanden werden müssen (Gunderson & Siegel, 2001). Eine weitere Erklärung könnte sein, dass die in den IDS-2 gestellten Aufgaben kulturell beeinflusst und für Fremdsprachige möglicherweise ungewohnt waren. Dies stimmt mit dem Befund überein, dass viele Aufgaben nebst sprachlichen und kulturellen Einflüssen auch von aufgabenspezifischen Vorerfahrungen abhängig sind (Grigorenko & Sternberg, 1998). Dahingegen könnte ein Grund für fehlende Effekte in *Unpassende Bilder ausschließen* darin liegen, dass—ähnlich wie bei *Figuren wiedererkennen* und *Rotierte Figuren wiedererkennen*—im Sinne des DT ein Lerneffekt von *Matrizen ergänzen* zu *Unpassende Bilder ausschließen* stattgefunden hat (vgl. Guthke & Wiedl, 1996).

In Bezug auf die Vergleiche zwischen Mono- und Bilingualen zeigte die Gruppe der Bilingualen zwar—nebst der allgemeinen Intelligenz—leicht tiefere Werte im *Langzeitgedächtnis* sowie in *Denken Abstrakt* und *Denken Verbal*, die Unterschiede wurden jedoch nicht signifikant.

Die vorliegenden Befunde deuten darauf hin, dass die IDS-2 hinsichtlich der Intelligenzeinschätzung teilweise sprachgebunden sind, dass aber durchaus Faktoren und Untertests identifiziert werden können, die nur eine geringe oder überhaupt keine Sprachgebundenheit aufweisen. Während die ersten zwei Faktoren *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* sowie deren Untertests keine Unterschiede zwischen monolingualen, zwei- und fremdsprachigen Kindern und Jugendlichen zeigen, ergeben sich in den anderen fünf Faktoren sowie einigen Untertests zum *Kurzzeitgedächtnis* und *Langzeitgedächtnis* sowie zum *Denken Abstrakt* und *Denken Verbal* Unterschiede zuungunsten von Fremdsprachigen. *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* erfassen basale Intelligenzaspekte wie visuelle Diskriminations- und Verarbeitungsfähigkeiten—die anderen IDS-2-Faktoren sind hingegen stärker sprachgebunden und weisen eine höhere Aufgaben- und Instruktionskomplexität auf (vgl. Grob & Hagmann-von Arx, 2018). Dies wirkt sich insbesondere auf die Leistungen von Fremdsprachigen aus, welche im Gegensatz zu Mono- und Bilingualen weniger mit der deutschen Sprache in Berührung kommen und dementsprechend womöglich über tiefere Sprachkenntnisse verfügen. Aufgrund des sprachgebundenen und kristallinen Inhalts von *Denken Verbal* wird in diesem Faktor und den zugehörigen Untertests am meisten Varianz durch die Gruppenzugehörigkeit erklärt und die Unterschiede zuungunsten der Fremdsprachigen sind am größten.

Die IDS-2 weisen im Vergleich zu den IDS-P und IDS eine höhere Sprachgebundenheit auf, was zu größeren Mittelwertdifferenzen zuungunsten von fremdsprachigen Personen führt. Setzt man die IDS-2 weiter in Relation zur WISC-IV, werden einige Parallelen, aber auch Unterschiede deutlich, wobei die Befunde aufgrund der unterschiedlichen Stichprobenzusammensetzungen schwer zu vergleichen sind: Während die Studien zur WISC-IV (Daseking et al., 2008; Hagmann-von Arx et al., 2013) Unterschiede zwischen Deutsch- und Fremdsprachigen—letztere definiert als Personen mit Migrationshintergrund mit Zweitsprache Deutsch—untersuchten, unterscheidet die vorliegende Studie zwischen Monolingualen, Bilingualen und Fremdsprachigen. Bezüglich der allgemeinen Intelligenz setzt sich diejenige der IDS-2 ähnlich derjenigen der WISC-IV sowohl aus fluiden als auch kristallinen Aspekten zusammen, was in beiden Testverfahren zu tieferen Mittelwerten für Fremdsprachige führt (WISC-IV: kleine Effekte; IDS-2: kleine, nicht signifikante Effekte für Bilinguale und mittlere Effekte für Fremdsprachige). Während die WISC-IV

keine Unterschiede in der *Verarbeitungsgeschwindigkeit* und im *Arbeitsgedächtnis*, jedoch kleine Unterschiede zuungunsten von Fremdsprachigen in der fluiden (*Wahrnehmungsgebundenes Logisches Denken*) und kristallinen (*Sprachverständnis*) Intelligenz findet, kommen in den IDS-2 ebenfalls keine Unterschiede in der *Verarbeitungsgeschwindigkeit*, jedoch ebenfalls Unterschiede zuungunsten von Fremdsprachigen in der fluiden (*Denken Abstrakt*) und kristallinen (*Denken Verbal*) Intelligenz zu Tage; weiter finden sich in den IDS-2 Unterschiede sowohl im phonologischen als auch räumlich-visuellen *Kurzzeitgedächtnis*. Während die Effekte für Fremdsprachige im kleinen bis großen Bereich liegen, sind diejenigen für Bilinguale maximal im kleinen, jedoch nicht signifikanten Bereich. Folglich empfiehlt sich für die Praxis, sich der teilweisen Sprachgebundenheit der IDS-2 und damit zusammenhängend einer möglichen Benachteiligung durch Fremdsprachigkeit bewusst zu sein: *Verarbeitung Visuell* und *Verarbeitungsgeschwindigkeit* weisen mit 0 bis 1 % erklärter Varianz die niedrigste Sprachgebundenheit auf, wohingegen *Kurzzeitgedächtnis Auditiv*, *Kurzzeitgedächtnis Räumlich-Visuell*, *Langzeitgedächtnis* sowie *Denken Abstrakt* mit 1 bis 7 % erklärter Varianz stärker sprachgebunden sind und *Denken Verbal* mit 12 bis 15 % erklärter Varianz die größte Sprachgebundenheit zeigt. Die IDS-2 erlauben nebst einer interindividuellen Auswertung die Betrachtung einer Profilanalyse mit spezifischen intraindividuellen Stärken und Schwächen. Sollte eine (große) Diskrepanz auf intraindividuellere Ebene vorliegen, kann diese womöglich auf die Sprachgebundenheit dieser Faktoren sowie der zugehörigen Untertests zurückgeführt werden. In diesem Falle sollte die allgemeine Intelligenz aufgrund des kristallinen Anteils mit Vorsicht interpretiert werden.

Stärken der vorliegenden Studie stellen die große Stichprobe sowie die große Altersspanne dar. Weiter wurden sowohl Fremd- als auch Zweisprachige im Vergleich mit monolingual Deutschsprachigen untersucht. Ebenfalls ist positiv hervorzuheben, dass für die Überprüfung der Gruppenunterschiede jeweils eine Paarung nach Alter, Geschlecht und mütterlichem Ausbildungsstand stattgefunden hat, um möglichst vergleichbare Gruppen zu gewährleisten. Als Einschränkung ist anzumerken, dass keine Angaben vorliegen, in welchem Alter die Fremd- und Zweisprachigen mit dem Erlernen von Deutsch begonnen haben, respektive wie gut ihre Deutschkenntnisse zum Zeitpunkt der Testdurchführung waren. Die hier untersuchten Stichproben weisen somit eine gewisse Heterogenität auf, was die Vergleichbarkeit innerhalb dieser Gruppen erschwert. Nichtsdestotrotz spiegeln die vielen verschiedenen zu Deutsch fremden Muttersprachen die sprachliche und ethnische Vielfalt in der Schweiz und in Deutschland wider (vgl. Bundesamt für Statistik, 2016; Statistisches Bundesamt, 2017). Eine weitere Limitation ist, dass keine vorgängige Überprüfung der Messinvarianz über die Gruppen hinweg stattgefunden hat.

Zukünftige Forschung sollte Leistungsunterschiede differenzierter in weiteren sprachlichen Subgruppen untersuchen, beispielsweise könnten verschiedene Fremdsprachen einzeln analysiert werden. Weiterführende Studien sollten zudem Differenzen zwischen Monolingualen, Bilingualen und Fremdsprachigen in den weiteren Funktionsbereichen der IDS-2 analysieren, um die Befundlage zur differentiellen Validität dieses Testverfahrens zu erweitern. Ebenfalls könnte untersucht werden, ob sich Unterschiede über verschiedene Altersgruppen hinweg unterscheiden. Da sprachliche Kompetenzen für erfolgreiche Bildungs- und Lernprozesse unerlässlich sind (Baumert et al., 2002), manifestieren sich sprachliche Defizite langfristig auch in nichtsprachlichen Leistungsbereichen und der Unterschied von Fremd- zu Deutschsprachigen wird mit zunehmendem Alter größer (Daseking et al., 2008).

Zusammenfassend lässt sich festhalten, dass je nach Sprachgebundenheit der Untertests und Faktoren der IDS-2 Mittelwertdifferenzen zwischen den untersuchten Stichproben auftreten. Die Ergebnisse in der visuellen Verarbeitung und Verarbeitungsgeschwindigkeit sind vergleichbar zwischen den Gruppen. Unterschiede aufgrund der Gruppenzugehörigkeit zeigen sich in der allgemeinen Intelligenz, im räumlich-visuellen und auditiven Kurzzeitgedächtnis, im Langzeitgedächtnis sowie im abstrakten und verbalen Denken, wobei diese für die allgemeine Intelligenz mit 7 bis 10

% und für verbales Denken mit 15 % erklärter Varianz am grössten ausfallen. Auf Stichprobenebene erzielen Fremdsprachige im räumlich-visuellen und auditiven Kurzzeitgedächtnis, im Langzeitgedächtnis sowie im abstrakten Denken tiefere Werte als Monolinguale, und in der allgemeinen Intelligenz sowie im verbalen Denken tiefere Werte als Mono- und Bilinguale. Entsprechend sollen in der psychologischen Praxis der sprachliche Hintergrund eines Kindes und Jugendlichen und damit zusammenhängend mögliche Benachteiligungen für Personen mit nicht respektive nicht ausschließlich deutscher Muttersprache berücksichtigt werden. Die Betrachtung einer Profilanalyse ermöglicht es, Leistungsdiscrepanzen zwischen weniger sprachgebundenen sowie sprachgebundenen Skalen zu identifizieren und vor dem Hintergrund der Mehrsprachigkeit zu interpretieren.

Elektronische Supplemente (ESM)

Das elektronische Supplement ist mit der Online-Version dieses Artikels verfügbar unter [Link wird vom Verlag eingefügt]

ESM. Korrelationen

Literatur

- Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., ... Weiss, M. (Hrsg.). (2002). *Pisa 2000—Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.
- Bundesamt für Statistik (2016). *Erste Ergebnisse der Erhebung zur Sprache, Religion und Kultur 2014, Teil Sprache*. Verfügbar unter: <https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.assetdetail.1000398.html>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NY: Erlbaum.
- Daseking, M., Lipsius, M., Petermann, F. & Waldmann, H.-C. (2008). Intelligenz und kulturelle Einflüsse. Differenzen im Intelligenzprofil bei Kindern mit Migrationshintergrund: Befunde zum HAWIK-IV. *Kindheit und Entwicklung*, 17, 76-89. <https://doi.org/10.1026/0942-5403.17.2.76>
- Daseking, M. & Petermann, F. (2015). Nonverbale Intelligenzdiagnostik: Sprachfreie Erhebung kognitiver Fähigkeiten und Prävention von Entwicklungsrisiken. *Gesundheitswesen*, 77, 791-792. <https://doi.org/10.1055/s-0035-1564145>
- Deary, I. J. (2009). Introduction to the special issue on cognitive epidemiology. *Intelligence*, 37, 517-519. <https://doi.org/10.1016/j.intell.2009.05.001>
- Flanagan, D. P. & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York, NY: Wiley.
- Goldstein, S., Princiotta, D. & Naglieri, J. A. (Hrsg.). (2015). *Handbook of intelligence. Evolutionary theory, historical perspective, and current concepts*. New York, NY: Springer. <https://doi.org/10.1007/978-1-4939-1562-0>
- Grigorenko, E. L. & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124, 75-111. <https://doi.org/10.1037/0033-2909.124.1.75>
- Grob, A. & Hagmann-von Arx, P. (2018). *Intelligence and Development Scale-2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche*. Bern: Hogrefe.
- Grob, A., Meyer, C. S. & Hagmann-von Arx, P. (2013). *Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5-10 Jahren* (2. überarb. Aufl.). Bern: Hans Huber.
- Grob, A., Reimann, G., Gut, J. & Frischknecht, M.-C. (2013). *Intelligence and Development Scales-Preschool (IDS-P). Intelligenz- und Entwicklungsskalen für das Vorschulalter*. Bern: Hans Huber.
- Gunderson, L. & Siegel, L. S. (2001). The evils of the use of IQ tests to define learning disabilities in first- and second-language learners. *The Reading Teacher*, 55, 48-55.

- Guthke, J. & Wiedl, K. H. (1996). *Dynamisches Testen*. Göttingen: Hogrefe.
- Hagmann-von Arx, P., Meyer, C. S. & Grob, A. (2008). Intelligenz- und Entwicklungsdiagnostik im deutschen Sprachraum. *Kindheit und Entwicklung*, 17, 232-242. <https://doi.org/10.1026/0942-5403.17.4.232>
- Hagmann-von Arx, P., Petermann, F. & Grob, A. (2013). Konvergente und diskriminante Validität der WISC-IV und der Intelligence and Development Scales (IDS) bei Kindern mit Migrationshintergrund. *Diagnostica*, 59, 170-182. <https://doi.org/10.1026/0012-1924/a000091>
- Hessels, M. G. (1997). Low IQ but high learning potential: Why Zeyneb and Moussa do not belong in special education. *Educational and Child Psychology*, 14, 121-136.
- Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 43, 1-28.
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383-386.
- Janke, N., Daseking, M. & Petermann, F. (2008). Intelligenzdiagnostik im Kindergartenalter—Ein Beitrag zur Validierung des SON-R 2½-7. *Diagnostica*, 54, 174-183. <https://doi.org/10.1026/0012-1924.54.4.174>
- McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.) (pp. 136-182). New York, NY: Guilford Press.
- Petermann, F. & Petermann, U. (2011). *Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV)—Deutsche Version*. Frankfurt a. M.: Pearson Assessment.
- R Core Team (2016). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.
- Reich, H., Roth, H.-J., Dirim, I., Jörgensen, J. N., List, G., Neumann, U., ... Wurnig, V. (Hrsg.). (2002). *Spracherwerb zweisprachig aufwachsender Kinder und Jugendlicher. Ein Überblick über den Stand der nationalen und internationalen Forschung*. Hamburg: Behörde für Bildung und Sport.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F. & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137. <https://doi.org/10.1016/j.intell.2015.09.002>
- Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin: Springer. <https://doi.org/10.1007/978-3-642-17001-0>
- Schölmerich, A., Leyendecker, B., Citlak, B., Caspar, U. & Jäkel, J. (2008). Assessment of migrant and minority children. *Zeitschrift für Psychologie*, 216, 187-194. <https://doi.org/10.1027/0044-3409.216.3.187>
- Statistisches Bundesamt (2017). *Migration & Integration*. Verfügbar unter: <https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/MigrationIntegration/MigrationIntegration.html#Tabellen>
- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, 35, 401-426. <https://doi.org/10.1016/j.intell.2006.09.004>
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2007). *Snijders Oomen nonverbaler Intelligenztest 2½-7 (SON-R 2½-7)*. Göttingen: Hogrefe.
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2012). *Snijders Oomen nonverbaler Intelligenztest 6-40 (SON-R 6-40)*. Göttingen: Hogrefe.
- Toussaint, A., Heinze, L., Lipsius, M. & Petermann, F. (2012). Zur Aussagekraft des SON-R 6-40 bei Kindern mit Hörbeeinträchtigung und Kindern mit Migrationshintergrund. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 61, 108-121. <https://doi.org/10.13109/prkk.2012.61.2.108>
- Weiss, L. G., Harris, J. G., Prifitera, A., Courville, T., Rolfhus, E., Saklofske, D. H. & Holdnack, J. A. (2006). WISC-IV interpretation in societal context. The essentials and beyond. In L. G. Weiss, D. H. Saklofske, A. Prifitera & J. A. Holdnack (Eds.), *WISC-IV advanced clinical interpretation* (pp. 1-49). London: Elsevier.

Tabelle 1. Deskriptive Statistik für monolinguale, bilinguale und fremdsprachige Kinder und Jugendliche in der Intelligenz der IDS-2

Variable	Monolinguale			Bilinguale			Fremdsprachige		
	α^d	<i>M</i>	<i>SD</i>	α^d	<i>M</i>	<i>SD</i>	α^d	<i>M</i>	<i>SD</i>
Allgemeine Intelligenz									
IQ-Profil ^a	.98	101.23	16.41	.99	96.30	17.00	.98	90.49	16.58
IQ ^b	.97	101.20	16.75	.98	96.74	16.71	.97	90.06	16.97
IQ-Screening ^c	.95	101.60	17.36	.96	96.73	15.59	.95	88.42	16.65
Faktoren									
Verarbeitung Visuell	.95	99.31	15.69	.96	97.86	16.96	.96	95.46	17.66
Verarbeitungsgeschwindigkeit	.98	99.05	16.97	.98	96.24	17.95	.97	97.36	18.46
Kurzzeitgedächtnis Auditiv	.93	102.06	15.75	.94	99.05	16.10	.94	94.11	16.84
Kurzzeitgedächtnis Räumlich-Visuell	.92	99.90	14.60	.92	98.59	14.53	.92	94.71	14.09
Langzeitgedächtnis	.95	101.19	15.94	.96	95.26	15.47	.95	92.75	17.63
Denken Abstrakt	.94	101.16	16.92	.94	97.93	15.72	.94	93.66	16.38
Denken Verbal	.96	102.10	16.12	.97	96.21	15.53	.96	85.63	16.61
Untertests									
Figuren nachlegen	.91	10.10	3.03	.93	9.86	3.42	.92	9.54	3.39
Plättchen legen	.94	10.09	3.37	.95	9.86	3.09	.95	9.39	3.50
Zwei Merkmale durchstreichen	.97	9.91	3.37	.97	9.45	3.21	.93	9.63	3.53
Kästchen durchstreichen	.96	9.96	3.27	.96	9.46	3.58	.96	9.65	3.41
Zahlen- und Buchstabenreihen nachsprechen	.87	10.76	3.24	.89	10.32	3.25	.89	9.34	3.29
Gemischte Zahlen- und Buchstabenreihen nachsprechen	.85	10.90	3.08	.86	10.33	3.04	.87	9.44	3.40
Figuren wiedererkennen	.87	10.44	2.92	.86	10.15	2.87	.87	9.25	2.83
Rotierte Figuren wiedererkennen	.86	10.08	2.89	.86	9.86	2.81	.87	9.52	2.95
Geschichte nachherzählen	.93	10.40	3.30	.94	9.36	3.18	.93	8.22	3.60
Bild beschreiben	.91	10.52	3.01	.92	9.59	3.25	.91	9.87	3.24
Matrizen ergänzen	.92	10.63	3.34	.92	10.00	3.08	.92	9.26	3.13
Unpassende Bilder ausschließen	.88	10.45	3.25	.89	10.05	3.08	.90	9.34	3.26
Kategorien nennen	.93	10.69	3.25	.94	9.77	3.06	.93	7.76	3.56
Gegenteile nennen	.91	10.96	3.21	.93	9.75	3.28	.92	7.95	3.27

Anmerkungen: *n* = 143; Allgemeine Intelligenz und Faktoren: *M* = 100, *SD* = 15; Untertests: *M* = 10, *SD* = 3.^a Alle Untertests.^b Untertests Figuren nachlegen, Zwei Merkmale durchstreichen, Zahlen- und Buchstabenreihen nachsprechen, Figuren wiedererkennen, Geschichte nachherzählen, Matrizen ergänzen und Kategorien nennen.^c Untertests Matrizen ergänzen und Kategorien nennen.^d Cronbachs α .

Tabelle 2. Mittelwertvergleiche zwischen monolingualen, bilingualen und fremdsprachigen Kindern und Jugendlichen in der Intelligenz der IDS-2

Variable	Globaleffekte										Kontraste									
	Monolinguale – Fremdsprachige										Monolinguale – Bilinguale									
	<i>F</i>	<i>p_{HI}</i>	η^2	ΔM	<i>t</i>	<i>p_{HI}</i>	<i>d</i>	ΔM	<i>t</i>	<i>p_{HI}</i>	<i>d</i>	ΔM	<i>t</i>	<i>p_{HI}</i>	Bilinguale – Fremdsprachige					
Allgemeine Intelligenz	IQ-Profil ^a	14.55	<.001	.07	10.74	5.38	<.001	0.65	4.93	2.45	.48	0.30	5.81	2.92	<.001	0.35				
Faktoren	IQ ^b	15.64	<.001	.07	11.14	5.55	<.001	0.66	4.46	2.20	.64	0.27	6.68	3.33	<.001	0.40				
	IQ-Screening ^c	22.94	<.001	.10	13.18	6.69	<.001	0.77	4.87	2.46	.47	0.30	8.31	4.23	<.001	0.52				
	Verarbeitung Visuell	1.88	.51	.01	3.85	1.92	.65	0.23	1.45	0.72	.65	0.09	2.40	1.20	.65	0.14				
Untertests	Verarbeitungsgeschwindigkeit	0.88	.51	.00	1.69	0.79	.65	0.10	2.81	1.31	.65	0.16	-1.12	-0.53	.65	-0.06				
	Kurzzeitgedächtnis Auditiv	8.59	<.001	.04	7.95	4.10	<.001	0.49	3.01	1.54	.65	0.19	4.94	2.56	.38	0.30				
	Kurzzeitgedächtnis Räumlich-Visuell	4.96	.07	.02	5.19	3.02	.12	0.36	1.31	0.75	.65	0.09	3.88	2.26	.62	0.27				
	Langzeitgedächtnis	9.84	<.001	.04	8.44	4.33	<.001	0.50	5.93	3.03	.12	0.38	2.51	1.29	.65	0.15				
	Denken Abstrakt	7.46	.01	.03	7.50	3.85	.01	0.45	3.23	1.65	.65	0.20	4.27	2.19	.65	0.27				
	Denken Verbal	37.99	<.001	.15	16.47	8.59	<.001	1.01	5.89	3.05	.11	0.37	10.58	5.53	<.001	0.66				
	Figuren nachlegen	1.04	.51	.00	0.56	1.44	.65	0.17	0.24	0.62	.65	0.07	0.32	0.82	.65	0.09				
	Plättchen legen	1.63	.51	.01	0.70	1.76	.65	0.20	0.23	0.56	.65	0.07	0.47	1.20	.65	0.14				
	Zwei Merkmale durchstreichen	0.68	.51	.00	0.28	0.71	.65	0.08	0.46	1.16	.65	0.14	-0.18	-0.46	.65	-0.05				
	Kästchen durchstreichen	0.75	.51	.00	0.31	0.75	.65	0.09	0.50	1.22	.65	0.15	-0.19	-0.47	.65	-0.05				
Untertests	Zahlen- und Buchstabenreihen nachspr.	6.98	.01	.03	1.42	3.64	.02	0.43	0.44	1.13	.65	0.14	0.98	2.52	.42	0.30				
	Gem. Zahlen- und Buchstabenreihen nachspr.	7.52	.01	.03	1.46	3.84	.01	0.45	0.57	1.49	.65	0.19	0.89	2.35	.56	0.28				
	Figuren wiedererkennen	6.58	.02	.03	1.19	3.47	.03	0.41	0.29	0.84	.65	0.10	0.90	2.63	.33	0.32				
	Rotierte Figuren wiedererkennen	1.33	.51	.01	0.56	1.62	.65	0.19	0.22	0.62	.65	0.08	0.34	0.99	.65	0.12				
	Geschichte nachherzählen	14.77	<.001	.07	2.18	5.43	<.001	0.63	1.04	2.58	.37	0.32	1.14	2.85	.18	0.34				
	Bild beschreiben	3.13	.36	.01	0.65	1.72	.65	0.21	0.93	2.44	.50	0.30	-0.28	-0.73	.65	-0.09				
	Matrizen ergänzen	6.52	.02	.03	1.37	3.61	.02	0.42	0.63	1.65	.65	0.20	0.74	1.96	.65	0.24				
	Unpassende Bilder ausschließen	4.37	.12	.02	1.11	2.91	.15	0.34	0.40	1.05	.65	0.13	0.71	1.86	.65	0.22				
	Kategorien nennen	29.34	<.001	.12	2.93	7.47	<.001	0.86	0.92	2.33	.58	0.29	2.01	5.16	<.001	0.61				
	Gegenteile nennen	30.63	<.001	.13	3.01	7.77	<.001	0.93	1.21	3.11	.09	0.37	1.80	4.65	<.001	0.55				

Anmerkungen: *n* = 143; *p_{HI}* = *p*-Wert korrigiert nach Hommel; ΔM = Mittelwertdifferenz.^a Alle Untertests.^b Untertests Figuren nachlegen, Zwei Merkmale durchstreichen, Zahlen- und Buchstabenreihen nachspr., Figuren wiedererkennen, Geschichte nachherzählen, Matrizen ergänzen und Kategorien nennen.^c Untertests Matrizen ergänzen und Kategorien nennen.

ESM. Korrelationen

Tabelle 1. Partialkorrelationen für monolinguale, bilinguale und fremdsprachige Kinder und Jugendliche in der Intelligenz der IDS-2

Var-	IQ-P	IQ	IQ-S	VV	VG	KA	KRV	LG	DA	DV	FN	PL	ZMD	KD	ZBN	GZBN	FW	RFW	GN	BB	ME	UBA	KN	GEN
IQ-P	.96	.87	.83	.68	.72	.66	.68	.69	.78	.78	.73	.45	.66	.65	.63	.62	.59	.61	.69	.49	.73	.61	.68	.73
IQ-S	.86/87	.87	.83	.63	.66	.66	.63	.65	.75	.79	.73	.34	.66	.54	.65	.60	.60	.51	.70	.40	.77	.52	.72	.70
VV	.83/84	.86/87	.83	.48	.49	.49	.43	.55	.80	.84	.58	.24	.44	.42	.49	.44	.38	.39	.56	.37	.86	.51	.85	.64
VG	.72/71	.69/65	.61/60	.45	.45	.37	.47	.33	.48	.40	.79	.84	.42	.42	.35	.37	.41	.41	.38	.20	.47	.35	.37	.36
KA	.72/72	.70/71	.52/58	.48/44	.37/46	.32	.44	.44	.52	.45	.46	.30	.91	.90	.32	.31	.35	.42	.39	.36	.50	.39	.35	.47
KRV	.72/62	.70/60	.53/39	.43/38	.41/38	.39/34	.43/45	.49/38	.48	.52	.39	.47	.32	.37	.46	.46	.89	.88	.32	.22	.39	.38	.28	.46
LG	.71/71	.65/67	.56/57	.38/35	.38/35	.41/39	.43/45	.49/38	.48	.52	.36	.19	.38	.40	.26	.27	.24	.32	.87	.84	.46	.37	.47	.46
DA	.77/73	.71/69	.79/75	.57/55	.47/55	.47/42	.59/41	.47/38	.49/41	.56	.54	.17	.39	.40	.42	.41	.36	.39	.48	.35	.87	.85	.50	.51
DV	.77/78	.75/75	.78/80	.46/52	.50/44	.40/43	.42/38	.38/36	.55/56	.46/51	.54	.35	.30	.27	.20	.23	.26	.30	.58	.30	.54	.43	.90	.89
FN	.68/68	.72/70	.58/61	.86/85	.49/41	.40/43	.42/38	.38/36	.55/56	.46/51	.54	.35	.30	.27	.20	.23	.26	.30	.58	.30	.54	.43	.90	.89
PL	.53/55	.43/43	.44/43	.83/86	.33/35	.27/29	.30/28	.25/27	.42/41	.30/40	.43/47	.33	.30	.27	.20	.23	.26	.30	.58	.30	.54	.43	.90	.89
ZMD	.69/65	.71/66	.51/51	.43/35	.90/92	.39/46	.41/36	.41/31	.47/49	.49/37	.46/43	.32/36	.66/68	.64	.26	.77	.36	.41	.43	.41	.22	.27	.21	.15
KD	.63/68	.58/64	.45/45	.47/45	.92/91	.29/38	.34/34	.33/41	.40/52	.44/43	.46/43	.32/36	.66/68	.64	.26	.77	.36	.41	.43	.41	.22	.27	.21	.15
ZBN	.61/67	.62/69	.42/47	.34/37	.31/47	.94/94	.36/29	.33/41	.38/40	.47/49	.35/41	.22/25	.32/46	.26/40	.26	.77	.36	.41	.43	.41	.22	.27	.21	.15
GZBN	.72/69	.68/65	.55/46	.41/41	.37/42	.94/94	.36/29	.33/41	.38/40	.47/49	.35/41	.22/25	.32/46	.26/40	.26	.77	.36	.41	.43	.41	.22	.27	.21	.15
FW	.60/47	.62/53	.42/28	.34/26	.37/32	.31/28	.88/83	.56/27	.49/26	.54/20	.32/34	.26/11	.40/30	.30/30	.32/28	.35/24	.33/29	.33/29	.39/37	.21/20	.44/51	.42/32	.61/62	.60
RFW	.66/55	.60/46	.52/37	.44/37	.34/31	.38/28	.87/83	.50/27	.56/43	.40/31	.42/35	.20/28	.46/30	.38/39	.35/46	.50/48	.55/38	.44/40	.31/52	.36/24	.49/37	.37/26	.61/64	.60
GN	.70/72	.73/72	.60/62	.38/36	.46/38	.46/49	.44/41	.80/89	.52/42	.66/67	.42/35	.21/18	.21/23	.17/31	.20/24	.28/27	.26/17	.36/24	.31/52	.36/24	.49/37	.37/26	.61/64	.60
BB	.45/50	.34/41	.31/35	.24/24	.21/30	.26/27	.36/24	.81/84	.25/22	.29/40	.19/26	.41/39	.40/52	.37/51	.29/35	.43/36	.42/32	.44/40	.31/52	.36/24	.49/37	.37/26	.61/64	.60
ME	.68/68	.69/70	.83/81	.56/53	.42/37	.38/37	.49/37	.37/34	.85/87	.40/39	.51/54	.29/32	.40/34	.30/41	.35/36	.42/32	.44/40	.44/40	.31/52	.36/24	.49/37	.37/26	.61/64	.60
UBA	.67/59	.61/50	.52/49	.42/43	.38/40	.42/35	.53/35	.42/32	.85/87	.44/34	.42/43	.29/34	.44/34	.37/41	.41/45	.49/41	.33/23	.34/30	.56/58	.21/36	.33/33	.37/29	.61/62	.60
KN	.71/71	.74/74	.83/84	.45/47	.44/41	.49/46	.40/30	.55/59	.46/39	.89/91	.45/47	.29/34	.44/34	.37/41	.41/45	.49/41	.33/23	.34/30	.56/58	.21/36	.33/33	.37/29	.61/62	.60
GEN	.67/69	.61/61	.57/59	.35/47	.47/41	.52/50	.35/27	.47/55	.41/37	.90/89	.35/44	.24/39	.44/36	.42/38	.45/46	.54/47	.27/13	.34/30	.56/58	.21/36	.33/33	.37/29	.61/62	.60

Anmerkungen. $n = 143$; IQ-P = IQ-Profil; IQ-S = IQ-Screening; VV = Verarbeitung Visuell; VG = Verarbeitung Visuell; KA = Kurzzeitgedächtnis; KRV = Kurzzeitgedächtnis Räumlich-Visuell; LG = Langzeitgedächtnis; DA = Denken Abstrakt; DV = Denken Verbal; FN = Figuren nachlegen; PL = Plättchen legen; ZMD = Zwei Merkmale durchstreichen; KD = Kästchen durchstreichen; ZBN = Zahlen- und Buchstabenreihen nachsprechen; GZBN = Gemischte Zahlen- und Buchstabenreihen nachsprechen; FW = Figuren wiedererkennen; RFW = Rotierte Figuren wiedererkennen; GN = Geschichte nachzählen; BB = Bild beschreiben; ME = Matrizen ergänzen; UBA = Unpassende Bilder ausschließen; KN = Kategorien nennen; GEN = Gegenteil nennen.^a Oberhalb der Diagonalen: Werte für Monolinguale; Unterhalb der Diagonalen: Werte für Bilinguale links und Werte für Fremdsprachige rechts vom Trennstrich; Korrelationen sind kontrolliert für Alter und Geschlecht und korrigiert nach Hommel; Angenommene höchste Korrelationen sind fett gedruckt; Signifikanz: $r \leq .19$ n.s.; $.20 \leq r \leq .29$, $p < .05$; $.30 \leq r \leq .37$, $p < .01$; $r \geq .38$, $p < .001$.

APPENDIX C: Studie 3

Gygi, J. T., Hagmann-von Arx, P., Schweizer, F. & Grob, A. (2017). The predictive validity of four intelligence tests for school grades: A small sample longitudinal study. *Frontiers in Psychology*, 8, 375. doi: 10.3389/fpsyg.2017.00375



The Predictive Validity of Four Intelligence Tests for School Grades: A Small Sample Longitudinal Study

Jasmin T. Gygi, Priska Hagmann-von Arx*, Florine Schweizer and Alexander Grob

Department of Psychology, University of Basel, Basel, Switzerland

OPEN ACCESS

Edited by:

Ann Dowker,
University of Oxford, UK

Reviewed by:

Chris Lange-Küttner,
London Metropolitan University, UK
Bert Jonsson,
Umeå University, Sweden

*Correspondence:

Priska Hagmann-von Arx
priska.hagmann@unibas.ch

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 03 October 2016

Accepted: 27 February 2017

Published: 13 March 2017

Citation:

Gygi JT, Hagmann-von Arx P,
Schweizer F and Grob A (2017) The
Predictive Validity of Four Intelligence
Tests for School Grades: A Small
Sample Longitudinal Study.
Front. Psychol. 8:375.
doi: 10.3389/fpsyg.2017.00375

Intelligence is considered the strongest single predictor of scholastic achievement. However, little is known regarding the predictive validity of well-established intelligence tests for school grades. We analyzed the predictive validity of four widely used intelligence tests in German-speaking countries: The Intelligence and Development Scales (IDS), the Reynolds Intellectual Assessment Scales (RIAS), the Snijders-Oomen Nonverbal Intelligence Test (SON-R 6-40), and the Wechsler Intelligence Scale for Children (WISC-IV), which were individually administered to 103 children ($M_{\text{age}} = 9.17$ years) enrolled in regular school. School grades were collected longitudinally after 3 years (averaged school grades, mathematics, and language) and were available for 54 children ($M_{\text{age}} = 11.77$ years). All four tests significantly predicted averaged school grades. Furthermore, the IDS and the RIAS predicted both mathematics and language, while the SON-R 6-40 predicted mathematics. The WISC-IV showed no significant association with longitudinal scholastic achievement when mathematics and language were analyzed separately. The results revealed the predictive validity of currently used intelligence tests for longitudinal scholastic achievement in German-speaking countries and support their use in psychological practice, in particular for predicting averaged school grades. However, this conclusion has to be considered as preliminary due to the small sample of children observed.

Keywords: validity, scholastic achievement, IDS, RIAS, SON-R 6-40, WISC-IV

INTRODUCTION

The primary purpose of the first intelligence test (Binet and Simon, 1905) was to predict scholastic achievement in order to determine the best school setting for a child. Since the beginning of intelligence assessment, the predictive validity of intelligence test scores for scholastic achievement has been well studied. Cross-sectional and longitudinal studies indicated strong correlations, around $r = 0.40$ – 0.81 , between the two (e.g., Sternberg et al., 2001; Deary et al., 2007; Mackintosh, 2011).

The association between intelligence and scholastic achievement seems to be stronger when using standardized achievement tests compared to school grades (Sternberg et al., 2001; Rost, 2009). Standardized achievement tests represent achievement at only one point in time, whereas school grades represent achievement over a longer period and thus may also be influenced by other constructs such as self-control and motivation (Rost, 2009). However, school grades are crucial for children to be promoted to the next higher grade level as well as for further scholastic and occupational qualifications (Roth et al., 2015).

Focusing on school grades, a recent meta-analysis (Roth et al., 2015) found an observed correlation of $r = 0.44$ and an estimated true correlation (i.e., corrected for error of measurement and range restriction) of $\rho = 0.54$ between intelligence and school grades. Regarding subject domains, the correlations were highest and comparable for mathematics/science ($r = 0.42$, $\rho = 0.49$) and languages ($r = 0.36$, $\rho = 0.44$). The results furthermore revealed that correlations between intelligence and school grades in elementary school ($r = 0.40$, $\rho = 0.45$) tended to be weaker than in middle and high school ($r = 0.46$, $\rho = 0.54$ – 0.58), because intelligence deficits in elementary school may be compensated more easily through practice than in higher-grade levels, as the learning content is easier to understand. This result is in contrast to previous research (e.g., Sternberg et al., 2001), that identified stronger correlations between intelligence and scholastic achievement in elementary school than in higher-grade levels, because of growing range restrictions.

The meta-analysis performed by Roth et al. (2015) included studies conducted in different countries. In German-speaking countries, for example, the Culture Fair Test-20-Revision (Weiss, 2006), standardized in 2003, showed associations with school grades in mathematics/science ranging from $r = 0.26$ to 0.39 and in languages of $r = 0.23$. Further, the German Cognitive Ability Test – 4-12 – Revision (KFT 4-12+R; Heller and Perleth, 2000), standardized from 1995 to 1997, showed associations with school grades in mathematics/science ranging from $r = 0.17$ to 0.60 and in languages ranging from $r = 0.12$ to 0.14 . In another study, the KFT 4-12+R and the German version of the Wechsler Intelligence Scale for Children-III (Tewes et al., 1999), standardized from 1995 to 1998, predicted mathematics/science with $\beta = 0.54$ and language with $\beta = 0.52$ (Karbach et al., 2013). However, the meta-analysis did not include more recently standardized intelligence tests currently used in German-speaking countries.

Currently used intelligence tests in German-speaking countries (e.g., Hagmann-von Arx et al., 2015) include (a) the Intelligence and Development Scales (IDS; Grob et al., 2013), an intelligence test for children aged 5–10 years measuring in particular fluid intelligence; (b) the Reynolds Intellectual Assessment Scales (RIAS; Reynolds and Kamphaus, 2003; German version: Hagmann-von Arx and Grob, 2014), an intelligence test for individuals aged 3 to above 90 years that measures verbal and nonverbal intelligence, based on crystallized and fluid intelligence, respectively. A composite intelligence index can be computed from the values in verbal and nonverbal intelligence; (c) the Snijders-Oomen Nonverbal Intelligence Test Revised 6-40 (SON-R 6-40; Tellegen et al., 2012), a nonverbal intelligence test measuring fluid intelligence in individuals aged 6–40 years; and (d) the Wechsler Intelligence Scales for Children, Fourth Edition (WISC-IV; Wechsler, 2003; German version: Petermann and Petermann, 2011), an intelligence test used worldwide to measure general intelligence (Full-Scale IQ or FSIQ). Additionally, the WISC-IV provides four index scores: verbal comprehension reflecting the understanding of verbal concepts; perceptual reasoning measuring nonverbal perception and manipulation; working memory assessing attention and

working memory; and processing speed reflecting visuospatial speed of processing.

Nevertheless, little is known regarding the predictive validity of these intelligence tests for school grades. Especially for German-speaking countries and for studies independent of the standardization samples, there is a lack of literature analyzing predictive validity of these tests. For the IDS, Gut et al. (2013) analyzed the predictive validity of general intelligence in children aged 5–7 years from the standardization sample for concurrent ($n = 402$) and longitudinal ($n = 221$) scholastic achievement. Concurrent scholastic achievement was operationalized through parents' and teachers' ratings in mathematics, science, and language (German), which were averaged across subjects. Longitudinal scholastic achievement was based on averaged school grades in these subjects 3 years later. Results revealed medium to large effect sizes ($\beta = 0.30$ – 0.56) for the cross-sectional and a small effect size ($\beta = 0.21$) for the longitudinal association. These results replicate findings of a prior study conducted by Gut et al. (2012) showing that in an extended sample of 263 children aged 5–10 years, IDS general intelligence predicted school grades in mathematics, science, and language (German) 3 years later with medium effect sizes ($\beta = 0.28$ – 0.34). Both studies indicate small to moderate concurrent and predictive validity of the IDS general intelligence for averaged school grades.

For the German version of the RIAS, we found no studies on the predictive validity of intelligence indices on school grades. However, the Technical Manual of the English Version of the RIAS (Reynolds and Kamphaus, 2003) reports a cross-sectional validation study conducted with 78 children aged 3–16 years. Results revealed strong correlations between the composite intelligence index and a standardized achievement test in mathematics ($r = 0.67$) and language ($r = 0.64$), indicating strong concurrent validity of the RIAS composite intelligence index for standardized achievement tests.

For the SON-R 6-40, the Technical Manual of the German version (Tellegen et al., 2012) reports moderate to strong correlations between the test scores and concurrent school grades in mathematics ($r = 0.58$) and language ($r = 0.49$) for 182 elementary school children aged 6–11 years. These results indicate that nonverbal intelligence measured using the SON-R 6-40 shows moderate to strong concurrent validity for school grades.

For the German version of the WISC-IV there are, to our knowledge, no available studies on the predictive validity for school grades. For the English version of the WISC-IV, Glutting et al. (2006) studied the concurrent validity of general intelligence and its specific indices on a standardized academic achievement test in mathematics and reading with a sample of 498 individuals aged 6–16 years from the WISC-IV standardization sample. Results showed large effect sizes for the WISC-IV FSIQ (60% variance explained) but only small effect sizes for the specific indices (0–2% additional variance explained). The FSIQ predicted concurrent mathematics and reading equally well. Thus, results indicate that in particular the WISC-IV FSIQ is correlated with concurrent standardized academic achievement tests.

In sum, Roth et al.'s (2015) meta-analysis revealed that intelligence test scores correlate moderately to strongly with school grades in mathematics and language. However, there is only scarce evidence regarding the longitudinal prediction of school grades with currently used intelligence tests in German-speaking countries.

In the current study we analyzed the predictive power of the German versions of the IDS, the RIAS, the SON-R 6-40, and the WISC-IV for longitudinal school grades. We analyzed the general intelligence indices only, as the Technical Manuals (e.g., Reynolds and Kamphaus, 2003) as well as previous research (Glutting et al., 2006) do not recommend the use of specific indices for high-stakes decisions because of lowered reliability and validity compared to the general intelligence indices. On the basis of previous research (e.g., Glutting et al., 2006; Deary et al., 2007; Gut et al., 2012, 2013; Roth et al., 2015), we expected that the general intelligence indices would positively predict averaged school grades as well as school grades in mathematics and language (German) with medium to strong effect sizes.

MATERIALS AND METHODS

Participants

The sample consisted of 103 children aged 6–11 years ($M = 9.17$ years, $SD = 0.93$; 52% females, 48% males) enrolled in regular schools. All children took part in an intelligence assessment. Three years later, parents of 54 children aged 10–13 years ($M = 11.77$ years, $SD = 0.79$; 52% females, 48% males) provided information about their children's school grades in mathematics and language. Regarding parental education, 74% of the parents had a non-tertiary education and 26% had a tertiary education. This distribution indicates that parent's educational attainment in the present study is comparable with the general Swiss population (Swiss Federal Statistical Office, 2016). Post hoc power analysis using G*Power (Faul et al., 2007) indicated that with a chance of 80% and a 0.05 alpha level, the current study was sufficiently powered to detect medium effect sizes ($r = 0.30$; Cohen, 1988). The 54 children who participated in both study waves showed significantly higher intelligence scores in the RIAS composite intelligence index ($M = 103.24$, $SD = 8.29$) as well as in the WISC-IV FSIQ ($M = 107.39$, $SD = 10.58$) than the 49 children who did not participate in the second study wave (RIAS: $M = 98.98$, $SD = 9.19$, $F = 0.82$, $p < 0.01$; WISC-IV: $M = 102.12$, $SD = 12.78$, $F = 1.57$, $p < 0.05$). No differences were found for the IDS and the SON-R 6-40.

Measures Intelligence

To assess intelligence ($M = 100$, $SD = 15$), the IDS, RIAS, SON-R 6-40, and WISC-IV were administered. The IDS assesses general intelligence and five developmental domains (psychomotor skills, social-emotional competences, mathematics, language, and achievement motivation) in children aged 5–10 years. For the current study, only general intelligence was analyzed. IDS general intelligence consists of seven subtests (i.e., visual perception, selective attention, phonological memory, visual-spatial memory, auditory memory, abstract reasoning, figural

reasoning), which measure mainly fluid intelligence. The administration of IDS general intelligence takes about 45 min. The IDS was standardized from 2007 to 2008 in Austria, Germany, and Switzerland. Reliability for general intelligence is high with Cronbach's $\alpha = 0.92$.

The RIAS is an intelligence test for individuals aged 3 to above 90 years. It comprises four intelligence subtests (i.e., guess what, verbal reasoning, odd-item out, what's missing), which together constitute the composite intelligence index, CIX. The CIX can also be divided into two indices, represented by two of the four above mentioned subtests each: the Verbal Intelligence Index, VIX, representing crystallized intelligence, and the Nonverbal Intelligence Index, NIX, representing fluid intelligence. Two additional subtests can be administered measuring verbal and nonverbal memory resulting in a Composite Memory Index. The memory subtests are not entered in the CIX. The assessment of the RIAS CIX takes about 20–25 min. The German version of the RIAS was standardized from 2011 to 2012 in Germany and Switzerland. Reliability for the RIAS is high with Cronbach's $\alpha = 0.95$ for the CIX and $\alpha = 0.93$ – 0.94 for the VIX and NIX.

The SON-R 6-40 assesses nonverbal intelligence for individuals aged 6–40 years. It comprises four subtests (i.e., analogies, categories, mosaics, patterns) that primarily measure fluid intelligence. The administration of the SON-R 6-40 takes about 45–60 min. The German version of the SON-R 6-40 was standardized from 2009 to 2011 in Germany and the Netherlands. Reliability for the SON-R 6-40 is high with Cronbach's $\alpha = 0.95$.

The WISC-IV is an intelligence test measuring general intelligence for children aged 6–16 years. It includes 10 core subtests (i.e., similarities, vocabulary, comprehension, block design, picture concepts, matrix reasoning, digit span, letter-number sequencing, symbol search, coding) that constitute the FSIQ and four specific indices: the Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index. The administration of the WISC-IV core subtests takes about 60 min. The German version of the WISC-IV was standardized from 2005 to 2006 in Austria, Germany, and Switzerland. Reliability for the WISC-IV is high with $r = 0.97$ for the FSIQ and $r = 0.87$ – 0.94 for the specific intelligence indices.

School Grades

Three years after intelligence assessment, parents were asked to report on their child's school grades in mathematics and language (1 = poorest grade, 6 = best grade; grades 4–6 represent the passing range) based on the school records of the latest term (i.e., overall grades). In Switzerland, passing grades in both mathematics and language are crucial for a child to be promoted to the next higher grade level (Swiss Media Institute for Education and Culture, 2016). Thus, in line with previous research (e.g., Gut et al., 2013) school grades were additionally averaged across subjects to obtain a composite estimate of scholastic achievement.

Procedure

This study was carried out in accordance with the recommendations of the Ethics Committee of Basel, Switzerland

and with the Declaration of Helsinki. Parents gave written informed consent prior to participation in the study, and assent was obtained from the children. Children were recruited from elementary schools in the German-speaking part of Switzerland in 2011. Trained study personnel administered the tests at school on regular school days. Each child was individually administered the four intelligence tests (IDS, RIAS, SON-R 6-40, WISC-IV) in counterbalanced order. Three appointments were required, each about 2 h, including breaks (one test session for the IDS, one test session for the WISC-IV, and one test session for the RIAS and SON-R 6-40). The sample sizes for each intelligence test vary somewhat, as a few children could attend only two testing appointments ($n_{IDS} = 103$, $n_{RIAS} = 102$, $n_{SON-R6-40} = 101$, $n_{WISC-IV} = 103$). After the study was completed, the parents received a written report on their child's performance in each intelligence test. Three years later, parents were contacted again and asked to provide information about their child's school grades. Two families had moved and could not be reached; in all, 54 parents returned the requested information (resulting in a response rate of 53%).

Data Analyses

All analyses were conducted using SPSS 23.0. Because of the small sample size and because some of the variables showed deviations from normality (see skewness and kurtosis in **Table 1**), we used bootstrap procedures (Efron, 1979; Chernick, 2008). Bias-corrected 95% confidence intervals (BC 95%-CI) were computed based on 5,000 random samples. A result was considered to be significant when the confidence interval did not include zero.

To analyze the predictive validity of each intelligence test, separate regression analyses for each predictor (i.e., general intelligence indices) and outcome variable (i.e., child's school grades) were conducted. All variables entered into the regression analyses were z-standardized. Few children were identified as outliers with scores more than two standard deviations from the mean ($n_{IDS} = 2$, $n_{RIAS} = 1$, $n_{SON-R6-40} = 2$, $n_{WISC-IV} = 2$), and for this reason these scores were truncated to $z \pm 2$. In the following analyses, we controlled for variables that showed correlations with the outcome variables to some extent (i.e., sex, age; see **Table 2**).

RESULTS

Table 1 gives an overview of the descriptive statistics of the current sample. The mean scores of the intelligence tests were somewhat higher than in the standardization samples ($M = 100$), and the standard deviations were somewhat lower than in the standardization samples ($SD = 15$). The range of school grades is narrow (4–6) and reflects grades in the passing range. Correlations among all variables are displayed in **Table 2**. The general intelligence indices of all four tests correlated highly with each other ($r = 0.63$ – 0.80 , $p < 0.001$).

Table 3 presents the results of the prediction of longitudinal school grades by intelligence (model a: IDS, model b: RIAS, model c: SON-R 6-40, model d: WISC-IV) while controlling for children's sex and age. When school grades were averaged across subjects, the general intelligence indices of all four intelligence

tests predicted scholastic achievement. Estimates ranged from Estimate = 0.126 (SE = 0.054, BC 95%-CI = [0.019, 0.225]) for SON-R 6-40 to Estimate = 0.175 (SE = 0.059, BC 95%-CI = [0.056, 0.293]) for IDS.

Regarding mathematics, results show that IDS (Estimate = 0.376, SE = 0.143, BC 95%-CI = [0.085, 0.677]), RIAS (Estimate = 0.283, SE = 0.142, BC 95%-CI = [0.023, 0.544]), and SON-R 6-40 (Estimate = 0.275, SE = 0.137, BC 95%-CI = [0.007, 0.519]) were significant predictors, whereas WISC-IV was not significantly related to school grades in mathematics. Regarding language, IDS (Estimate = 0.318, SE = 0.130, BC 95%-CI = [0.066, 0.573]) and RIAS (Estimate = 0.376, SE = 0.153, BC 95%-CI = [0.092, 0.699]) significantly predicted school grades. The general intelligence indices of SON-R 6-40 and WISC-IV did not predict school grades in language.

DISCUSSION

Our main goal was to assess the longitudinal predictive validity of four intelligence tests currently used in German-speaking countries for children's school grades in mathematics, language, and averaged across subjects.

The general intelligence indices of all four intelligence tests showed significant predictive validity for averaged school grades three years after intelligence assessment, which is in line with previous studies, showing that intelligence is a positive predictor of scholastic achievement (Deary et al., 2007; Gut et al., 2012, 2013; Roth et al., 2015). Therefore, our results support the use of the general intelligence indices of the IDS, RIAS, SON-R 6-40, and WISC-IV in order to make predictions of a child's averaged school grades.

Regarding the prediction of mathematics, IDS, RIAS, and SON-R 6-40 were significantly associated with school grades. IDS general intelligence includes four out of seven subtests that tax phonological and visual-spatial working memory (i.e., selective attention, phonological memory, visual-spatial memory, auditory memory). Previous research revealed that both aspects of working memory are associated with mathematics (e.g., Dehn, 2008; Raghubar et al., 2010). The RIAS includes two out of four subtests taxing visual-spatial abilities, while the SON-R 6-40 assesses intelligence through subtests measuring primarily visual-spatial abilities. Previous literature found visual-spatial abilities to be moderately associated with mathematics (e.g., Wai et al., 2009; Verdine et al., 2014). The WISC-IV did not significantly predict school grades in mathematics, although the small effect sizes were positive; however, this contradicts the results of previous studies that found general intelligence to be a moderate to strong predictor of school grades in mathematics (Roth et al., 2015). The WISC-IV includes perceptual reasoning and phonological working memory as two out of four specific indices, and thus measure visual-spatial abilities and working memory to a lesser extent. This might have weakened the relation between these general intelligence indices and school grades in mathematics, as visual-spatial abilities (e.g., Verdine et al., 2014; Wai et al., 2009) and working memory (Dehn, 2008; Raghubar et al., 2010) were found to be predictors of mathematics. Thus,

TABLE 1 | Descriptive statistics for control variables, the intelligence scores of the IDS, RIAS, SON-R 6-40, WISC-IV, and school grades in study waves 1 and 2.

Variable	Study wave 1 (N = 103)					Study wave 2 (N = 54)				
	Mean (SD)	N (%)	Range	Skewness	Kurtosis	Mean (SD)	N (%)	Range	Skewness	Kurtosis
Sex										
Female		54 (52)					28 (52)			
Male		49 (48)					26 (48)			
Age at Study wave 1 (years)	9.18 (0.93)		6.71–11.18	−0.19	0.26	9.04 (0.96)		6.71–11.18	−0.08	−0.14
Age at Study wave 2 (years)						11.77 (0.80)		10.25–13.60	0.43	−0.51
Parental education										
Non-tertiary education		74 (72)					40 (74)			
Tertiary education		29 (28)					14 (26)			
IDS	103.54 (10.64)		78–134	0.15	0.41	105.19 (10.74)		79–134	0.45	0.70
RIAS	101.45 (8.99)		81–128	0.29	0.39	103.65 (8.23)		87–128	0.84	0.67
SON-R 6-40	103.27 (11.32)		79–139	0.38	0.50	103.67 (11.55)		82–139	0.62	0.61
WISC-IV	104.88 (11.91)		73–133	0.21	0.07	107.39 (10.58)		84–133	0.27	0.28
Averaged school grade						5.16 (0.42)		4–6	−0.39	−0.20
Mathematics school grade						5.15 (0.48)		4–6	−0.25	−0.71
Language school grade						5.16 (0.43)		4–6	−0.31	−0.08

IDS, Intelligence and Development Scales; RIAS, Reynolds Intellectual Assessment Scales; SON-R 6-40, Snijders-Oomen Nonverbal Intelligence Test Revised; WISC-IV, Wechsler Intelligence Scale for Children—Fourth Edition.

TABLE 2 | Correlations among control variables, the intelligence scores of the IDS, RIAS, SON-R 6-40, WISC-IV, and school grades.

Variable	N	1	2	3	4	5	6	7	8	9	10
1 Sex (0 = males, 1 = females)	54	1									
2 Age at Study wave 1	54	−0.09	1								
3 Age at Study wave 2	54	−0.09	0.79***	1							
4 Parental education (0 = non-tertiary, 1 = tertiary)	54	−0.02	−0.14	−0.06	1						
5 IDS	54	0.16	−0.34*	−0.24	0.13	1					
6 RIAS	54	−0.03	−0.31*	−0.13	0.15	0.75***	1				
7 SON-R 6-40	52	0.05	−0.16	−0.18	0.11	0.80***	0.63***	1			
8 WISC-IV	54	0.09	−0.27*	−0.07	−0.02	0.77***	0.67***	0.63***	1		
9 Averaged school grade	54	0.24	−0.30*	−0.37**	0.10	0.47**	0.34**	0.36*	0.30*	1	
10 Mathematics school grade	54	0.14	−0.25	−0.28*	0.13	0.43**	0.26	0.32*	0.25	0.93***	1
11 Language school grade	53	0.32*	−0.30*	−0.41**	0.06	0.44**	0.36**	0.32*	0.30*	0.91***	0.68***

Correlations were calculated for individuals who participated at Study wave 2. IDS, Intelligence and Development Scales; RIAS, Reynolds Intellectual Assessment Scales; SON-R 6-40, Snijders-Oomen Nonverbal Intelligence Test Revised; WISC-IV, Wechsler Intelligence Scale for Children—Fourth Edition.

* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

in the IDS, phonological and visual-spatial working memory capacity, and in the RIAS and SON-R 6-40, visual-spatial abilities are considered more important parts of intelligence compared to the other intelligence tests. Therefore, it might be plausible that in particular IDS, RIAS, and SON-R 6-40 were significantly associated with school grades in mathematics.

Regarding the prediction of language, the general intelligence indices of the IDS and RIAS were significantly associated with school grades in language. The association between the IDS and language is in line with studies revealing a moderate to strong relationship between working memory and language (e.g., Dehn, 2008). The association between the RIAS and language might be explained through the high requirements of verbal

abilities and verbal reasoning in two out of four RIAS subtests. The other general intelligence indices showed no significant associations with language, although the small effect sizes were positive; however, this result contradicts the findings of previous studies that found general intelligence to be a moderate to strong predictor of school grades in language (Roth et al., 2015). In contrast to the RIAS, the SON-R 6-40 focuses only on nonverbal intelligence. Furthermore and in contrast to the IDS, the SON-R 6-40 does not include subtests taxing working memory, which is considered as being associated with language (e.g., Dehn, 2008). The WISC-IV includes a specific index for verbal comprehension and working memory. However, it might be possible that these two out of four specific indices were not sufficient to

TABLE 3 | Regression analyses for the intelligence scores of the IDS, RIAS, SON-R 6-40, and WISC-IV predicting longitudinal school grades.

Variables	Averaged school grade				Mathematics				Language			
	Estimate	SE	BC 95%-CI	R ²	Estimate	SE	BC 95%-CI	R ²	Estimate	SE	BC 95%-CI	R ²
Model a: IDS												
Step 1				0.180*				0.092				0.251*
Sex (0 = males, 1 = females)	0.086	0.052	[-0.010, 0.176]		0.112	0.122	[-0.118, 0.324]		0.264	0.110	[0.055, 0.473]	
Age	-0.165	0.060	[-0.281, -0.019]		-0.281	0.140	[-0.540, 0.063]		-0.386	0.115	[-0.622, -0.137]	
Step 2 ^a				0.318*				0.220**				0.341**
IDS	0.175	0.059	[0.056, 0.293]		0.376	0.143	[0.085, 0.677]		0.318	0.130	[0.066, 0.573]	
Model b: RIAS												
Step 1				0.180*				0.092				0.251*
Sex (0 = males, 1 = females)	0.086	0.052	[-0.007, 0.177]		0.112	0.120	[-0.127, 0.338]		0.264	0.112	[0.059, 0.476]	
Age	-0.165	0.060	[-0.266, -0.043]		-0.281	0.139	[-0.518, 0.031]		-0.386	0.117	[-0.621, -0.150]	
Step 2 ^a				0.275*				0.148				0.347**
RIAS	0.165	0.066	[0.031, 0.298]		0.283	0.142	[0.023, 0.544]		0.376	0.153	[0.092, 0.699]	
Model c: SON-R 6-40												
Step 1				0.179*				0.093				0.251*
Sex (0 = males, 1 = females)	0.069	0.053	[-0.030, 0.164]		0.086	0.126	[-0.151, 0.310]		0.220	0.112	[0.013, 0.418]	
Age	-0.170	0.060	[-0.284, -0.031]		-0.283	0.141	[-0.554, 0.072]		-0.406	0.114	[-0.625, -0.165]	
Step 2 ^a				0.262*				0.168*				0.301*
SON-R 6-40	0.126	0.054	[0.019, 0.225]		0.275	0.137	[0.007, 0.519]		0.217	0.116	[-0.020, 0.443]	
Model d: WISC-IV												
Step 1				0.180*				0.092				0.251*
Sex (0 = males, 1 = females)	0.086	0.051	[-0.013, 0.186]		0.112	0.120	[-0.114, 0.328]		0.264	0.114	[0.048, 0.476]	
Age	-0.165	0.060	[-0.270, -0.029]		-0.281	0.140	[-0.531, 0.067]		-0.386	0.115	[-0.615, -0.146]	
Step 2 ^a				0.247*				0.140				0.304**
WISC-IV	0.128	0.063	[0.001, 0.266]		0.242	0.157	[-0.088, 0.590]		0.260	0.137	[-0.001, 0.553]	

IDS, Intelligence and Development Scales; RIAS, Reynolds Intellectual Assessment Scales; SON-R 6-40, Sijders-Corren Nonverbal Intelligence Test Revised; WISC-IV, Wechsler Intelligence Scale for Children—Fourth Edition. BC 95%-CI, bias-corrected 95% bootstrap confidence intervals.

^aControlled for variables in step 1.

* $p < 0.01$, ** $p < 0.001$.

significantly explain variance in school grades in language in the present study. Thus, the different subtests underlying the general intelligence indices of the IDS, RIAS, SON-R 6-40, and WISC-IV may be responsible for the different associations with school grades in mathematics and language. However, future studies with larger sample sizes have to be conducted to analyze this assumption.

It is notable that the effect sizes in the present study were in the small to moderate range and thus somewhat lower than we expected based on the meta-analytical results by Roth et al. (2015). However, in a single study, the expected effects may be smaller, as seen, for example, in Gut et al. (2012, 2013), for several reasons. In the present study, for instance, the analyzed sample showed slightly higher general intelligence scores than the population and had a narrow range in school grades, which were all in the passing range. This might have led to range restrictions, which may have weakened the correlations between intelligence and school grades in the present study (Sternberg et al., 2001; Roth et al., 2015). Also, the present study analyzed the predictive validity of intelligence tests on school grades for elementary school children. According to Roth et al. (2015), lower intelligence scores in elementary school might be better compensated with practice than in higher-grade levels, which may also have led to smaller effect sizes in the present study.

Regarding the control variables, in the present study sex was significantly related to school grades such that girls achieved higher scores in language than boys. This result is in accordance with a recent meta-analysis showing that females have advantages in school marks which are largest for languages courses (Voyer and Voyer, 2014). Furthermore, age was negatively related to school grades such that older children achieved lower school grades than younger children. A possible explanation for this relation may be that age is related to pubertal status and that a more advanced physical pubertal status is related to lower achievement motivation (i.e., academic self-efficacy and valuing of school) that in turn is related to lower achievement (Martin and Steinbeck, 2017).

In sum, our results indicate that the general intelligence indices of the German versions of the IDS, RIAS, SON-R 6-40, and the WISC-IV significantly predicted averaged school grades over three years. Furthermore, the IDS and RIAS were positively associated with longitudinal mathematics and language school grades, while SON-R 6-40 was a predictor of mathematics school grades. Thus, our results provide evidence for predictive validity of these intelligence tests (Neukrug and Fawcett, 2015).

The current study has strengths and limitations. It is a strength that we analyzed four intelligence tests currently used in German-speaking countries, as there is a paucity of information regarding their predictive validity for school grades. Furthermore, we assessed intelligence three years prior to school grades being inquired and could therefore analyze their predictive validity longitudinally. This is especially relevant when practitioners use intelligence scores in order to predict future scholastic achievement. Finally, we measured a child's scholastic achievement in mathematics and language using school grades, which reflect a child's performance and effort over an extended period of time and which are crucial for further scholastic and

occupational qualifications (Roth et al., 2015). However, the association between intelligence and scholastic achievement may vary with different operationalization of scholastic achievement. In order to avoid potential errors in parental reports, future studies analyzing currently used intelligence tests in German-speaking countries might also consider achievement tests, which measure specific scholastic abilities at a specific point in time (Rost, 2009), as well as official school records obtained directly from schools. Moreover, as school grades are considered as indicators of achievement over a longer time period, they may be influenced not only by intelligence but also by other constructs (Rost, 2009). Therefore, future studies might also consider noncognitive factors that additionally predict scholastic achievement, such as school engagement (Reyes et al., 2012), motivation (Steinmayr and Spinath, 2009), self-control (Duckworth et al., 2012), personality (Poropat, 2009), and social-emotional competencies (Gut et al., 2012).

Furthermore, the current study had a high drop-out rate (although comparable to that of the studies conducted by Gut et al., 2012, 2013) for the longitudinal information on the child's school grades. This led to a small sample size at Study wave 2. The statistical power of the present study was sufficient to detect expected moderate associations, but there was not enough statistical power to detect weak associations between intelligence and school grades, as discussed above. Furthermore, the present study examined typically developing children enrolled in regular school with slightly higher intelligence. Thus, the conclusions based on the current study cannot be generalized to children with special needs or with different intelligence levels. To examine the predictive validity of the present intelligence tests, future studies are required with larger sample sizes and including children with different levels of intelligence (e.g., children with intellectual disabilities) or special needs as seen in the studies of Canivez et al. (2014), Mayes and Calhoun (2007), as well as Nelson and Canivez (2012). Because of these limitations, conclusions from the current study have to be considered as preliminary.

In conclusion, general intelligence measured with the German version of the IDS, RIAS, SON-R 6-40, and WISC-IV was a positive predictor of averaged school grades in the current longitudinal study. These results support the use of the four intelligence tests for issues raised in psychological practice and reveal their predictive validity on longitudinal scholastic achievement in typically developing school-aged children with slightly higher intelligence. Furthermore, the IDS and RIAS could predict both school grades in mathematics and language, while the SON-R 6-40 could predict school grades in mathematics. These results suggest that school grades in mathematics and language can be predicted by intelligence tests depending on their composition of subtests (e.g., working memory, verbal abilities, visual-spatial abilities). Thus, in psychological practice, examiners have to consider the variety of subtests included in a particular intelligence test when making specific predictions of mathematics and language. More studies analyzing larger samples as well as children with different levels of intelligence or special needs are required to replicate and generalize the findings of the current study.

AUTHOR CONTRIBUTIONS

JG and PH contributed to the study design, acquisition, analysis, and interpretation of data. Drafted and revised the manuscript, gave final approval, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. FS contributed to the study design and acquisition of data. Revised the manuscript, gave final approval, and agrees to be accountable for all aspects of the work in ensuring that questions related to the

accuracy or integrity of any part of the work are appropriately investigated and resolved. AG contributed to the study design and interpretation of data. Revised the manuscript, gave final approval, and agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

We thank Anita Todd and Laura Wiles for proofreading.

REFERENCES

- Binet, A., and Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychol.* 11, 191–336.
- Canivez, G. L., Watkins, M. W., James, T., Good, R., and James, K. (2014). Incremental validity of WISC-IV^{UK} factor index scores with a referred Irish sample: predicting performance on the WIAT-II^{UK}. *Br. J. Educ. Psychol.* 84, 667–684. doi: 10.1111/bjep.12056
- Chernick, M. R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Hoboken, NJ: John Wiley & Sons.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.
- Deary, I. J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence* 35, 13–21. doi: 10.1016/j.intell.2006.02.001
- Dehn, M. J. (2008). *Working Memory and Academic Learning. Assessment and Intervention*. Hoboken, NJ: Wiley.
- Duckworth, A. L., Quinn, P. D., and Tsukayama, E. (2012). What no child left behind leaves behind: the roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *J. Educ. Psychol.* 104, 439–451. doi: 10.1037/a0026280
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552
- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–179. doi: 10.3758/BF03193146
- Glutting, J. J., Watkins, M. W., Konold, T. R., and McDermott, P. A. (2006). Distinctions without a difference: the utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *J. Spec. Educ.* 40, 103–114. doi: 10.1177/00224669060400020101
- Grob, A., Meyer, C. S., and Hagemann-von Arx, P. (2013). *Intelligence and Development Scales (IDS)*, 2nd Edn. Bern: Hans Huber.
- Gut, J., Reimann, G., and Grob, A. (2012). Cognitive, language, mathematical, and socioemotional skills as predictors of subsequent school performance. *Z. Pädagogische Psychol.* 26, 213–200. doi: 10.1024/1010-0652/a000070
- Gut, J., Reimann, G., and Grob, A. (2013). A contextualized view on long-term predictors of academic performance. *J. Educ. Psychol.* 105, 436–443. doi: 10.1037/a0031503
- Hagemann-von Arx, P., Gauck, L., and Grob, A. (2015). "Intelligence assessment," in *Handbook of School Psychology*, eds K. Seifried, S. Drewes, and M. Hasselhorn (Stuttgart: Kohlhammer), 128–139.
- Hagemann-von Arx, P., and Grob, A. (2014). *Reynolds Intellectual Assessment Scales (RIAS)*. Bern: Hans Huber.
- Heller, K. A., and Perleth, C. (2000). *German Cognitive Ability Test – 4-12 – Revision*. Göttingen: Hogrefe.
- Karbach, J., Gottschling, J., Spengler, M., Hegewald, K., and Spinath, F. M. (2013). Parental involvement and general cognitive ability as predictors of domain-specific academic achievement in early adolescence. *Learn. Instr.* 23, 43–51. doi: 10.1016/j.learninstruc.2012.09.004
- Mackintosh, N. J. (2011). *IQ and Human Intelligence*, 2nd Edn. Oxford: Oxford University Press.
- Martin, A. J., and Steinbeck, K. (2017). The role of puberty in students' academic motivation and achievement. *Learn. Individ. Differ.* 53, 37–46. doi: 10.1016/j.lindif.2016.11.003
- Mayes, S. D., and Calhoun, S. L. (2007). Wechsler intelligence scale for children-third and -fourth edition predictors of academic achievement in children with attention-deficit/hyperactivity disorder. *Sch. Psychol. Q.* 22, 234–249. doi: 10.1037/1045-3830.22.2.234
- Nelson, J. M., and Canivez, G. L. (2012). Examination of the structural, convergent, and incremental validity of the Reynolds Intellectual Assessment Scales (RIAS) with a clinical sample. *Psychol. Assess.* 24, 129–140. doi: 10.1037/a0024878
- Neukrug, E. S., and Fawcett, R. C. (2015). *The Essentials of Testing and Assessment: A Practical Guide for Counselors, Social Workers, and Psychologists*, 3rd Edn. Stamford, CT: Cengage Learning.
- Petermann, F., and Petermann, U. (2011). *WISC-IV: Wechsler Intelligence Scale for Children—Fourth Edition (German version)*. Frankfurt: Pearson Assessment.
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* 135, 322–338. doi: 10.1037/a0014996
- Raghubar, K. P., Barnes, M. A., and Hecht, S. A. (2010). Working memory and mathematics: a review of developmental, individual difference, and cognitive approaches. *Learn. Individ. Differ.* 20, 110–122. doi: 10.1016/j.lindif.2009.10.005
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., and Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *J. Educ. Psychol.* 104, 700–712. doi: 10.1037/a0027268
- Reynolds, C. R., and Kamphaus, R. W. (2003). *RIAS, Reynolds Intellectual Assessment Scales*. Lutz, FL: PAR.
- Rost, D. H. (2009). *Intelligence: Facts and myths*. Weinheim: Beltz.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., and Spinath, F. M. (2015). Intelligence and school grades: a meta-analysis. *Intelligence* 53, 118–137. doi: 10.1016/j.intell.2015.09.002
- Swiss Federal Statistical Office (2016). *A Portrait of Switzerland. Results from the 2010-2014 Censuses*. Available online at: <https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/bildungsstand-kompetenzen/bevoelkerung.html>
- Swiss Media Institute for Education and Culture, (2016). *Compulsory Education*. Available online at: <http://swisseducation.educa.ch/en/compulsory-education-0>
- Steinmayr, R., and Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learn. Individ. Differ.* 19, 80–90. doi: 10.1016/j.lindif.2008.05.004
- Sternberg, R. J., Grigorenko, E., and Bundy, D. A. (2001). The predictive value of IQ. *Merrill Palmer Q.* 47, 1–41. doi: 10.1353/mpq.2001.0005
- Tellegen, P. J., Laros, J. A., and Petermann, F. (2012). *Snijders-Oomen Nonverbal Intelligence Test (SON-R 6-40)*. Göttingen: Hogrefe.
- Tewes, U., Rossmann, P., and Schallberger, U. (1999). *HAWIK-III. Hamburg-Wechsler-Intelligenztest für Kinder – Dritte Auflage. Übersetzung und Adaptation der WISC-III Wechsler Intelligence Scale for Children – Third Edition von David Wechsler*. Bern: Huber.
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Filipowicz, A. T., and Chang, A. (2014). Deconstructing building blocks: preschoolers' spatial assembly performance relates to early mathematical skills. *Child Dev.* 85, 1062–1076. doi: 10.1111/cdev.12165

- Voyer, D., and Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychol. Bull.* 140, 1174–1204. doi: 10.1037/a0036620
- Wai, J., Lubinski, D., and Benbow, C. P. (2009). Spatial ability for STEM domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* 101, 817–835. doi: 10.1037/a0016127
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV) Administration and Scoring Manual*. San Antonio, TX: The Psychological Association.
- Weiss, R. H. (2006). *Culture Fair Intelligence Test Scale 2—Revision (CFT 20-R)*. Göttingen: Hogrefe.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Gygi, Hagmann-von Arx, Schweizer and Grob. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX D: Studie 4

Schweizer, F., Gygi, J. T., Weidmann, R. & Grob, A. The concurrent and predictive validity of intelligence, conscientiousness, and achievement motivation for scholastic achievement across childhood and adolescence. *Manuskript zur Publikation eingereicht bei Personality and Individual Differences.*

Stand: 13.08.2018

The concurrent and predictive validity of intelligence, conscientiousness, and achievement
motivation for scholastic achievement across childhood and adolescence

Florine Schweizer, Jasmin Tamara Gygi, Rebekka Weidmann, and Alexander Grob
University of Basel, Department of Psychology, Missionsstrasse 62, 4055 Basel, Switzerland

Corresponding author

Florine Schweizer

Address: University of Basel, Department of Psychology,

Missionsstrasse 62, 4055 Basel, Switzerland

E-mail: florine.schweizer@unibas.ch

Telephone: +41 (0)61 207 06 37

Declarations of interest

None

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

1 The concurrent and predictive validity of intelligence, conscientiousness, and achievement
2 motivation for scholastic achievement across childhood and adolescence

3

4 **Abstract**

5 The present study investigated the concurrent and predictive validity of intelligence,
6 conscientiousness, and achievement motivation for scholastic achievement. The sample
7 consisted of 301 Swiss children and adolescents who took part in an intelligence assessment
8 in 2007 (M_{age} 8.04 years, $SD = 1.61$) and provided information about conscientiousness,
9 different facets of achievement motivation, and school grades (i.e., language, math, and
10 average grade) 7 years later (M_{age} 15.45 years, $SD = 1.52$; 48% male, 52% female).
11 Intelligence was a longitudinal predictor and conscientiousness and achievement striving were
12 cross-sectional predictors of school grades. Additionally, inhibitory test anxiety was a
13 negative concurrent predictor of average and math grades. Results indicate that intelligence
14 plays an important role in scholastic achievement across childhood and adolescence, while
15 traits and motivation show incremental validity over and above intelligence. These findings
16 are discussed in light of neo-socioanalytic theory and current research results.

17

18 **Keywords**

19 Intelligence; conscientiousness; achievement motivation; scholastic achievement; school
20 grades

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

21 **1. Introduction**

22 Education is a key factor in a myriad of important life outcomes. Apart from affecting
23 individuals' health and longevity (Deary, 2009), education is crucial for current and future
24 academic and occupational success (Roth et al., 2015; Strenze, 2007). Moreover, education is
25 not only beneficial for individuals themselves but also positively associated with important
26 developmental outcomes for the next generation (Bradley & Corwyn, 2002). Consequently, it
27 is of great interest to investigate what personality characteristics have an impact on good
28 education. Our objective in the present study was to examine the concurrent and predictive
29 validity¹ of intelligence, conscientiousness, and motivation for scholastic achievement in
30 terms of language, math, and average (across the two subjects) grades in children and
31 adolescents.

32 According to Roberts and Wood's (2006) neo-socioanalytic theory, personality
33 consists of four domains—abilities, traits, motives/values, and narratives—that lead to
34 different outcomes, such as academic achievement. The domain of abilities refers to
35 intelligence (Roberts & Wood, 2006), describing the “can do” (Gottfredson, 2002, p. 37) or
36 the “ability to do academic work” (Di Domenico & Fournier, 2015, p. 156). Traits and
37 motives, in contrast, are often described as the “will do” (Gottfredson, 2003, p. 369). Traits
38 subsume enduring patterns of thoughts, feelings, and behaviors (Roberts & Wood, 2006). For
39 scholastic achievement, especially conscientiousness, that is, the “readiness to do academic
40 work” (Di Domenico & Fournier, 2015, p. 156), is crucial (e.g., Richardson, Abraham, &
41 Bond, 2012). Motives and values include all things that are desirable to a person (Roberts &
42 Wood, 2006), for example, the “willingness to do academic work” (Di Domenico & Fournier,
43 2015, p. 156). In scholastic contexts, especially achievement motivation is crucial (e.g.,

¹ Concurrent validity is given when personality characteristics can predict achievement cross-sectionally, whereas predictive validity applies when achievement can be predicted over time, i.e., longitudinally.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

44 Robbins et al., 2004). Finally, narratives focus on how people create their life story (Roberts
45 & Wood, 2006).

46 The present study focused on abilities, traits, and motives, as these three domains have
47 been consistently linked to academic achievement (e.g., Di Domenico & Fournier, 2015).
48 Here, we outline current evidence on intelligence, conscientiousness, and achievement
49 motivation and their contributions to scholastic achievement in terms of school grades.

50 Intelligence is a general mental capacity for reasoning, problem solving, planning, and
51 learning (Gottfredson, 1997). Meta-analyses of cross-sectional and longitudinal studies
52 reported corrected correlations of $\rho = .20$ (Richardson et al., 2012; postsecondary school
53 level) and $\rho = .54$ (Roth et al., 2015; elementary to secondary school level) with school
54 grades. With regard to specific subjects, correlations are strongest for math and science ($\rho =$
55 $.49$) and language ($\rho = .44$; Roth et al., 2015).

56 Even though intelligence accounts for up to a third of the variance in scholastic
57 achievement, a considerable amount remains unexplained. Traits could be an additional
58 characteristic to take into account. Among the Big Five factors (Costa & McCrae, 1992),
59 conscientiousness is considered the strongest predictor. It assesses the extent to which
60 individuals are self-disciplined, dutiful, and achievement oriented (Costa & McCrae, 1992).
61 Meta-analyses have reported corrected correlations of $\rho = .19$ (Richardson et al., 2012;
62 postsecondary school level), $\rho = .22$ (Poropat, 2009; elementary to postsecondary school
63 level), $\rho = .24$ (O'Connor & Paunonen, 2007; postsecondary school level), and $\rho = .50$
64 (Poropat, 2014; elementary school level) with average grades. In single studies, for specific
65 subjects, conscientiousness was associated with math, science, and language grades (e.g.,
66 Dumfart & Neubauer, 2016; Spinath, Freudenthaler, & Neubauer, 2010; $r = .11-.31$).

67 In addition to intelligence and traits, motivation has been associated with achievement.
68 Motivation is a broad and heterogeneous concept that encompasses different aspects of
69 personality, including beliefs, values, and goals (Eccles & Wigfield, 2002). Regarding school

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

70 grades, achievement motivation, that is, a person's perseverance and performance on
71 achievement tasks (Wigfield & Eccles, 2000), seems to be an additional relevant factor
72 (Robbins et al., 2004). Meta-analyses have reported corrected correlations of subordinate
73 achievement motivation of $\rho = .30$ (Robbins et al., 2004; postsecondary school level) with
74 average grade. On the facet level, particularly self-efficacy ($\rho = .31-.59$) and test anxiety ($\rho =$
75 $-.24$) correlated with average grade (Richardson et al., 2012; postsecondary school level).
76 Moreover, in single studies, the facet achievement striving² was associated with average,
77 math, and language grades (e.g., Steinmayr & Spinath, 2007; $r = .22-.37$).

78 Studies combining intelligence, conscientiousness, and/or achievement motivation in
79 predicting scholastic achievement usually identified intelligence as the strongest predictor (cf.
80 Roth et al., 2015). Nevertheless, some studies found conscientiousness (e.g., Di Fabio &
81 Busoni, 2007) and achievement motivation (e.g., Steinmayr & Spinath, 2009) to show
82 incremental validity over intelligence, and another study reported incremental validity of
83 achievement motivation over traits (e.g., Steinmayr & Spinath, 2007). However, research
84 simultaneously examining intelligence, conscientiousness, and achievement motivation to
85 predict their specific and incremental levels of variance in scholastic achievement is scarce.

86 Therefore, in the present study we investigated the concurrent and predictive validity
87 of intelligence, conscientiousness, and achievement motivation for school grades. More
88 specifically, we examined intelligence as a longitudinal predictor and conscientiousness and
89 achievement motivation as cross-sectional predictors. The current study extends previous
90 research in three ways. First, examining the incremental predictive power of the three
91 characteristics provides much-needed research on the simultaneous impact of intelligence,
92 conscientiousness, and achievement motivation on scholastic achievement. Second, many

² According to personality (e.g., Costa & McCrae, 1992) and motivation (e.g., Eccles & Wigfield, 2002) conceptualizations, achievement striving represents a facet of conscientiousness as well as of achievement motivation. As we did not consider conscientiousness on the facet level, we subsumed achievement striving under achievement motivation in the present study.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

93 studies of intelligence predicting school grades are either cross-sectional or longitudinal over
94 only a short time. We analyzed the predictive validity of intelligence over a longer time span,
95 (i.e., 7 years). Finally, as motivation is a heterogeneous concept, previous studies have
96 produced inconsistent results and are difficult to directly compare. Examination of
97 achievement motivation at the facet level provides a more fine-grained picture of what aspects
98 are crucial for scholastic achievement.

99 Considering the above-mentioned findings, we expected intelligence,
100 conscientiousness, and achievement motivation to show concurrent and predictive validity for
101 school grades, and we formulated the following hypotheses: (1) Intelligence positively
102 predicts average, language, and math grades over 7 years. (2) Conscientiousness positively
103 predicts average, language, and math grades cross-sectionally above and beyond intelligence.
104 (3) Achievement motivation predicts average, language, and math grades cross-sectionally,
105 above and beyond intelligence and conscientiousness. Specifically, achievement striving
106 positively predicts and inhibitory test anxiety negatively predicts average, language, and math
107 grades.

2. Method

2.1 Participants and Procedure

The sample consisted of 301 participants who initially took part in the standardization study ($N = 1,330$) of the Intelligence and Development Scales (IDS; Grob, Meyer, & Hagmann-von Arx, 2009) in the German-speaking part of Switzerland in 2007. Children aged 5–10 years ($M = 8.04$ years, $SD = 1.61$; 50% male) were recruited from kindergartens (preschools) and elementary schools in 2007. Seven years later, in 2014, a questionnaire was sent to 747 families in the standardization study to get information about the child's further development (e.g., conscientiousness, achievement motivation, and school grades). Ninety families could not be reached. From the remaining 657 families, a total of 301 parents and their adolescent child aged 12–18 years ($M = 15.45$ years, $SD = 1.52$; 48% male) returned the

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

questionnaire (46% response rate), after informed consent was obtained. Two hundred and twenty-three (74%) mothers had no postsecondary education and 69 (23%) had postsecondary education, which is representative of the Swiss population (Swiss Federal Statistical Office, 2018). The families who returned the questionnaire showed no differences regarding age, sex, maternal education, and intelligence compared to the families who did not return the questionnaire.

2.2 Measures

2.2.1 Intelligence. Intelligence was assessed in 2007 using the IDS (Grob et al., 2009). This test battery assesses general intelligence and five developmental domains. General intelligence consists of visual perception, selective attention, memory (phonological, visuospatial, and auditory), and reasoning (visual and conceptual), with a standardized mean of 100 ($SD = 15$).

2.2.2 Conscientiousness. Conscientiousness was assessed in 2014 using the German Five Factors Questionnaire for Children [Fünf-Faktoren-Fragebogen für Kinder] (FFFK; Asendorpf, 1998). This parental questionnaire assesses extraversion, emotional stability, agreeableness, conscientiousness, and culture, with eight items per factor. For the current study, conscientiousness (sum score) was analyzed. The items consist of bipolar adjectives and are answered on a 5-point Likert scale ranging from 1 to 5, for example, “My child is: sloppy vs. conscientious.”

2.2.3 Achievement motivation. Achievement motivation was assessed in 2014 using the German Achievement Motivation Questionnaire for 7th- to 13th-Grade Students [Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse] (FLM 7-13; Petermann & Winkel, 2007). This self-report questionnaire assesses five facets of achievement motivation with 30 items on a 5-point Likert scale ranging from 1 (*totally disagree*) to 5 (*totally agree*). The sum scores of three facets represent achievement-supporting attitudes: *achievement striving*, assessing hope for success, competency goals, and

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

145 performance goals (8 items; e.g., “At school, I want to be the best”), *endurance and effort*,
146 including hope for success, self-control, and discipline (6 items; e.g., “I take a lot of time
147 when preparing for tests”), and *activating test anxiety*, encompassing fear of failure and
148 activating anxiety before and during testing situations (4 items; e.g., “In tests, I am more
149 successful if I am a little nervous beforehand”). Two facets represent achievement-inhibiting
150 attitudes: *fear of success*, measuring fear of negative peer reactions as a result of one’s
151 success (8 items; e.g., “Students who make more effort than most other students make
152 themselves unpopular with others”), and *inhibitory test anxiety*, assessing fear of failure and
153 inhibiting anxiety before and during testing situations (4 items; e.g., “If I have to solve a
154 difficult task, I am afraid to fail”).

155 **2.2.4 Scholastic achievement.** Scholastic achievement was assessed in 2014 using
156 parent reports of children’s standardized school grades in language (German) and math,
157 according to the latest school reports. Additionally, grades were averaged across the two
158 subjects. The Swiss school system evaluates students’ performance with grades from 1
159 (lowest) to 6 (highest), with grades 4 through 6 being the passing range.

160 **2.3 Data analysis**

161 We first calculated descriptive statistics and Pearson correlations. Second, we
162 conducted three separate hierarchical regression analyses, one for each dependent variable
163 (i.e., average, language, and math grade). Because age, sex, and maternal education correlated
164 significantly with the dependent variables (see Table 1), they were entered as control
165 variables in Step 1 in the regression analyses. Intelligence was added in Step 2,
166 conscientiousness in Step 3, and achievement motivation in Step 4. We present standardized
167 correlation and regression coefficients ($> .10$ = small, $> .30$ = medium, and $> .50$ = large;
168 Cohen, 1988). Analyses were conducted using SPSS 22.

169 **3. Results**

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

170 Descriptive statistics and Pearson correlations are displayed in Table 1. Hierarchical
171 regression analyses are shown in Table 2. Concerning average grades, the control variables in
172 Step 1 explained 6% of the variance. Intelligence in Step 2 predicted average grades and
173 incrementally explained 7% of the variance. When conscientiousness was added in Step 3,
174 intelligence remained a significant predictor. Additionally, conscientiousness predicted
175 average grades beyond intelligence and incrementally explained 9% of the variance. When
176 achievement motivation was entered in Step 4, intelligence and conscientiousness remained
177 significant predictors. Additionally, achievement motivation predicted average grades beyond
178 intelligence and conscientiousness and incrementally explained 8% of the variance:
179 Achievement striving positively and inhibitory test anxiety negatively predicted average
180 grades; the other facets showed no significant association. The final model explained 30% of
181 the variance.

182 Regarding language grades, the control variables (Step 1) explained 10% of the
183 variance. Intelligence (Step 2) predicted language grades and incrementally explained 5% of
184 the variance. When conscientiousness was added (Step 3), intelligence remained a significant
185 predictor. Additionally, conscientiousness predicted language grades beyond intelligence and
186 incrementally explained 5% of the variance. When achievement motivation was entered (Step
187 4), intelligence and conscientiousness remained significant predictors. Additionally,
188 achievement motivation predicted language grades beyond intelligence and conscientiousness
189 and incrementally explained 5% of the variance: Achievement striving positively predicted
190 language grades, whereas the other facets showed no significant association. The final model
191 explained 25% of the variance.

192 Regarding math grades, the control variables explained 6% of the variance.
193 Intelligence (Step 2) predicted math grades and incrementally explained 5% of the variance.
194 Conscientiousness (Step 3) predicted math grades beyond intelligence and incrementally
195 explained 7% of the variance. Achievement motivation (Step 4) predicted math grades

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

196 beyond intelligence and conscientiousness and incrementally explained 7% of the variance:
197 Achievement striving positively and inhibitory test anxiety negatively predicted math grades.
198 The final model explained 25% of the variance.

199 **4. Discussion**

200 We sought to add to the knowledge of the concurrent, predictive, and incremental
201 validity of intelligence, conscientiousness, and achievement motivation for average, language,
202 and math grades in children and adolescents. The results confirm our first hypothesis, as
203 intelligence was a predictor of school grades over 7 years and explained between 5 and 7% of
204 the variance with small effect sizes. Our findings are in line with previous research on
205 intelligence as a longitudinal predictor of school grades (cf. Roth et al., 2015) and with other
206 studies of the IDS that found small to medium effects across 3 years (Gut, Reimann, & Grob,
207 2012, 2013; Gygi, Hagmann-von Arx, Schweizer, & Grob, 2017). Our effects are somewhat
208 weaker than those of other studies where intelligence was considered the strongest predictor
209 (cf. Roth et al., 2015), but they are explained by the quite long time period of 7 years, as
210 associations decline with increasing time intervals (Howitt & Cramer, 2008). As intelligence
211 eases comprehension and learning in school (Di Fabio & Busoni, 2007) and is a relatively
212 stable construct (Rost, 2009), it is reasonable that it accounts for variance in school grades
213 over the long term.

214 Concerning the second hypothesis, conscientiousness was a cross-sectional predictor
215 of school grades above and beyond intelligence and incrementally explained between 5 and
216 9% of the variance with small to medium effect sizes. Our results correspond to findings in
217 the literature showing that conscientiousness predicts school grades (cf. Poropat, 2009, 2014)
218 and accounts for incremental variance above and beyond intelligence (e.g., Di Fabio &
219 Busoni, 2007). Presumably, students who are more organized, focused, and perseverant
220 perform better scholastically (cf. Di Fabio & Busoni, 2007), as they, for instance, pay
221 attention in school and do their homework more diligently. Further, it might be that students'

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

222 conscientious behavior in the classroom influences teachers' ratings and, accordingly, school
223 grades (Rosander & Bäckström, 2014).

224 Concerning the third hypothesis, some of the achievement motivation facets cross-
225 sectionally and weakly predicted school grades above and beyond intelligence and
226 conscientiousness and incrementally explained between 5 and 8% of the variance:
227 Achievement striving was a predictor of average, language, and math grades, whereas
228 inhibitory test anxiety was a predictor of average and math grades but not language grades.
229 The other three facets showed no association with grades. This is in line with the literature
230 reporting that achievement motivation predicts school grades (cf. Robbins et al., 2004) and
231 explains incremental variance above and beyond intelligence and traits (e.g., Steinmayr &
232 Spinath, 2007, 2009), and that achievement striving (e.g., Steinmayr & Spinath, 2007) and
233 test anxiety (cf. Richardson et al., 2012) show associations with grades.

234 Students who are high in achievement striving intend to accomplish more than others
235 (Petermann & Winkel, 2007). They might, for instance, invest more time in doing their
236 homework or preparing for tests while others take time off. Further, as achievement striving
237 can also be understood as a facet of conscientiousness (Costa & McCrae, 1992), its
238 association with school grades seems reasonable. In the present study, however, achievement
239 striving and conscientiousness showed no significant correlation ($r = .07, p > .05$)—although
240 this might have been due to the different rating sources, that is, self-reports (achievement
241 striving) and parental reports (conscientiousness).

242 On the other hand, the association of inhibitory test anxiety with math but not
243 language grades could be explained by performance anxiety: Whereas math usually requires
244 specific answers that are either right or wrong, language subjects instead demand a conclusive
245 presentation and argumentation of a standpoint independent of its correctness. Possibly,
246 students with performance anxiety do not give answers in math tasks but dare to reply to
247 language tasks where they can describe their point of view. As a consequence, inhibitory

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

248 anxiety might be more predominant in math compared to language performance (Steinmayr &
249 Spinath, 2007). Further, it is possible that domain-specific self-concepts—ability self-
250 perceptions and expectancies for future success—might influence school grades (Steinmayr &
251 Spinath, 2007).

252 In sum, in the present study, the personality characteristics intelligence,
253 conscientiousness, and achievement motivation—together with age, sex, and maternal
254 education—explained up to 30% of the variance in school grades. When we compared the
255 unique variances the three predictors accounted for, the magnitudes were comparable. The
256 current results indicate that concerning scholastic achievement, all three characteristics should
257 be considered and fostered. Although its rank-order stability is fairly high, intelligence in
258 terms of mean levels can be raised by a stimulating environment (e.g., school attendance;
259 Rost, 2009; also see Ritchie & Tucker-Drob, 2018). Conscientiousness and achievement
260 motivation might be positively influenced more easily. Conscientious behaviors such as being
261 on time or keeping focused can be trained with little effort but can have a major impact
262 (Dumfart & Neubauer, 2016). Likewise, motivation depends on the domain as well as context
263 (cf. Eccles & Wigfield, 2002) and therefore can vary and change.

264 The study has several strengths: We employed a large sample size, examined cross-
265 sectional as well as longitudinal predictors of scholastic achievement, and considered abilities
266 as well as traits and motivation as predictors and three separate school grades as outcome
267 variables. As we used an intelligence test in addition to parent (i.e., conscientiousness) and
268 self-reports (i.e., achievement motivation), we reduced shared method variance by combining
269 different information sources.

270 One limitation is the rather high dropout rate. However, there were no differences
271 regarding age, sex, maternal education, and intelligence between study continuers and
272 dropouts. Another limitation concerns the reliability of the FLM 7-13, which was, in some
273 cases, only marginally sufficient. However, coefficients were comparable to those reported in

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

274 the standardization manual (Petermann & Winkel, 2007). Finally, our study included only
275 typically developing children and adolescents, and therefore, results cannot be generalized to
276 clinical samples or other age groups.

277 Even though the examined personality characteristics could account for up to one third
278 of variance in school grades, a considerable amount remains unexplained. Future research
279 might address additional predictors, such as actual behavior and interest in school, that affect
280 later life success (Spengler, Damian, & Roberts, 2018). In addition, the predictive validity of
281 intelligence over a period longer than 7 years as well as the predictive validity of
282 conscientiousness and achievement motivation should be analyzed to identify their long-term
283 impact. Finally, other dependent variables, such as other school grades and standardized
284 achievement tests, need to be further investigated. The bigger picture of school achievement
285 can be completed when more puzzle pieces are examined.

286 In conclusion, our study provides further insight into the concurrent, predictive, and
287 incremental validity of different personality characteristics for scholastic achievement in
288 terms of school grades. Intelligence was a longitudinal predictor and conscientiousness and
289 achievement striving cross-sectional predictors of school grades. In addition, inhibitory test
290 anxiety negatively and cross-sectionally predicted average and math grades. In view of this
291 and in accordance with Roberts and Wood's (2006) neo-socioanalytic theory, all the above-
292 mentioned layers—the *ability* (i.e., intelligence), the *readiness* (i.e., conscientiousness), and
293 the *willingness* (i.e., motivation; cf. Di Domenico & Fournier, 2015; Roberts & Wood,
294 2006)—are crucial aspects of scholastic achievement in terms of school grades in children
295 and adolescents.

296 **Funding**

297 This research did not receive any specific grant from funding agencies in the public,
298 commercial, or not-for-profit sectors.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

299

References

- 300 Asendorpf, J. B. (1998). *Fünf-Faktoren-Fragebogen für Kinder [Five Factors Questionnaire*
 301 *for Children]* (FFFK). Berlin, Germany: Humboldt University, Institute for Psychol-
 302 ogy.
- 303 Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development.
 304 *Annual Review of Psychology*, 53, 371-399.
 305 doi:10.1146/annurev.psych.53.100901.135233
- 306 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale,
 307 NY: Erlbaum.
- 308 Costa, P. T., & McCrae, R. R. (1992). *NEO-PI-R Professional Manual*. Odessa, FL: Psycho-
 309 logical Assessment Resources.
- 310 Deary, I. J. (2009). Introduction to the special issue on cognitive epidemiology. *Intelligence*,
 311 37, 517-519. doi:10.1016/j.intell.2009.05.001
- 312 Di Domenico, S. I., & Fournier, M. A. (2015). Able, ready, and willing: Examining the addi-
 313 tive and interactive effects of intelligence, conscientiousness, and autonomous moti-
 314 vation on undergraduate academic performance. *Learning and Individual Differ-*
 315 *ences*, 40, 156-162. doi:10.1016/j.lindif.2015.03.016
- 316 Di Fabio, A., & Busoni, L. (2007). Fluid intelligence, personality traits and scholastic suc-
 317 cess: Empirical evidence in a sample of Italian high school students. *Personality and*
 318 *Individual Differences*, 43, 2095-2104. doi:10.1016/j.paid.2007.06.025
- 319 Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness is the most powerful noncognitive
 320 predictor of school achievement in adolescents. *Journal of Individual Differences*,
 321 37, 8-15. doi:10.1027/1614-0001/a000182
- 322 Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review*
 323 *of Psychology*, 53, 109-132. doi:10.1146/annurev.psych.53.100901.135153
- 324 Gottfredson, L. S. (1997). Intelligence and social policy [Special issue]. *Intelligence*, 24.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

- 325 Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, 15,
 326 25-46. doi:10.1080/08959285.2002.9668082
- 327 Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence.
 328 *Intelligence*, 31, 343-397. doi:10.1016/S0160-2896(02)00085-5
- 329 Grob, A., Meyer, C. S., & Hagmann-von Arx, P. (2009). *Intelligence and Development Scales*
 330 *(IDS)*. Bern, Switzerland: Huber.
- 331 Gut, J., Reimann, G., & Grob, A. (2012). Kognitive, sprachliche, mathematische und sozial-
 332 emotionale Kompetenzen als Prädiktoren späterer schulischer Leistungen: Können
 333 die Leistungen eines Kindes in den IDS dessen Schulnoten drei Jahre später vorher-
 334 sagen? [Cognitive, linguistic, mathematical, and social-emotional skills as predictors
 335 for subsequent scholastic achievement: Is a child's performance on the IDS predic-
 336 tive for his/her school grades three years later?] *Zeitschrift für Pädagogische Psycho-*
 337 *logie*, 26, 213-220. doi:10.1024/1010-0652/a000070
- 338 Gut, J., Reimann, G., & Grob, A. (2013). A contextualized view on long-term predictors of
 339 academic performance. *Journal of Educational Psychology*, 105, 436-443.
 340 doi:10.1037/a0031503
- 341 Gygi, J. T., Hagmann-von Arx, P., Schweizer, F., & Grob, A. (2017). The predictive validity
 342 of four intelligence tests for school grades: A small sample longitudinal study. *Fron-*
 343 *tiers in Psychology*, 8, 375. doi:10.3389/fpsyg.2017.00375
- 344 Howitt, D., & Cramer, D. (2008). *Introduction to research methods in psychology* (2nd ed.).
 345 Harlow, England: Pearson Education Limited.
- 346 O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-
 347 secondary academic performance. *Personality and Individual Differences*, 43, 971-
 348 990. doi:10.1016/j.paid.2007.03.017

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

- 349 Petermann, F., & Winkel, S. (2007). *Fragebogen zur Leistungsmotivation für Schüler der 7.*
 350 *bis 13. Klasse [Achievement Motivation Questionnaire for 7th- to 13th-Grade Stu-*
 351 *dents] (FLM 7-13)*. Frankfurt, Germany: Harcourt Test Services.
- 352 Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic
 353 performance. *Psychological Bulletin*, 135, 322-338. doi:10.1037/a0014996
- 354 Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic perfor-
 355 mance in primary education. *British Journal of Educational Psychology*, 84, 239-
 356 252. doi:10.1111/bjep.12019
- 357 Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university stu-
 358 dents' academic performance: A systematic review and meta-analysis. *Psychological*
 359 *Bulletin*, 138, 353-387. doi:10.1037/a0026838
- 360 Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence?
 361 A meta-analysis. *Psychological Science*. Advance online publication.
 362 doi:10.1177/0956797618774253
- 363 Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psy-
 364 chosocial and study skill factors predict college outcomes? A meta-analysis. *Psycho-*
 365 *logical Bulletin*, 130, 261-288. doi:10.1037/0033-2909.130.2.261
- 366 Roberts, B. W., & Wood, D. (2006). Personality development in the context of the neo-
 367 socioanalytic model of personality. In D. Mroczek, & T. Little (Eds.), *Handbook of*
 368 *personality development* (pp. 11-39). Mahwah, NJ: Erlbaum.
 369 doi:10.4324/9781315805610.ch2
- 370 Rosander, P., & Bäckström, M. (2014). Personality traits measured at baseline can predict
 371 academic performance in upper secondary school three years later. *Scandinavian*
 372 *Journal of Psychology*, 55, 611-618. doi:10.1111/sjop.12165
- 373 Rost, D. H. (2009). *Intelligenz: Fakten und Mythen [Intelligence: Facts and myths]*. Wein-
 374 heim, Germany: Beltz.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

- 375 Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intel-
376 ligence and school grades: A meta-analysis. *Intelligence*, 53, 118-137.
377 doi:10.1016/j.intell.2015.09.002
- 378 Spengler, M., Damian, R. I., & Roberts, B. W. (2018). How you behave in school predicts life
379 success above and beyond family background, broad traits, and cognitive ability.
380 *Journal of Personality and Social Psychology*, 114, 620-636.
381 doi:10.1037/pspp0000185
- 382 Spinath, B., Freudenthaler, H., & Neubauer, A. C. (2010). Domain-specific school achieve-
383 ment in boys and girls as predicted by intelligence, personality and motivation. *Per-*
384 *sonality and Individual Differences*, 48, 481-486. doi:10.1016/j.paid.2009.11.028
- 385 Steinmayr, R., & Spinath, B. (2007). Predicting school achievement from motivation and per-
386 sonality. *Zeitschrift für Pädagogische Psychologie*, 21, 207-216. doi:10.1024/1010-
387 0652.21.3.207
- 388 Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school
389 achievement. *Learning and Individual Differences*, 19, 80-90.
390 doi:10.1016/j.lindif.2008.05.004
- 391 Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longi-
392 tudinal research. *Intelligence*, 35, 401-426. doi:10.1016/j.intell.2006.09.004
- 393 Swiss Federal Statistical Office. (2018). *Bildungsstand und Kompetenzen [Education and*
394 *skills]*. Retrieved from [https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-](https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/bildungsstand-kompetenzen.html)
395 [wissenschaft/bildungsstand-kompetenzen.html](https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/bildungsstand-kompetenzen.html)
- 396 Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation.
397 *Contemporary Educational Psychology*, 25, 68-81. doi:10.1006/ceps.1999.1015

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

1 **Acknowledgments**

2 We thank Priska Hagmann-von Arx for her comments on an earlier version of this
3 manuscript and Anita Todd for editing the manuscript.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

Table 1

Descriptive Statistics and Pearson Correlations

Variable	M	SD	Range	α	1	2	3	4	5	6	7	8	9	10	11	12
1 Age (years)	15.45	1.52	12.58–18.42													
2 Sex					.00											
3 Maternal education					-.16**	.00										
4 Intelligence	101.14	13.58	64–135	.92	-.09	.00	.28***									
5 Conscientiousness	29.93	5.37	6–40	.85	.13*	.24***	-.08	.01								
6 Achievement striving	25.86	5.47	9–39	.77	-.10	-.24***	.01	.08	.07							
7 Endurance and effort	19.57	4.10	8–28	.72	-.15*	.21***	-.05	-.05	.38***	.35***						
8 Fear of success	13.66	3.89	8–30	.67	-.04	-.03	-.03	.03	-.04	.08	-.12*					
9 Activating test anxiety	9.43	2.80	4–17	.68	.00	-.05	-.01	-.06	.07	.27***	.10	.22***				
10 Inhibitory test anxiety	11.17	3.19	4–20	.59	-.08	.13*	-.17**	-.13*	.12*	.05	.19**	.30***	.08			
11 Average grade	4.82	0.50	3.5–6		-.13*	.10	.15*	.29***	.30***	.27***	.14*	.07	.12	-.10		
12 Language grade	4.88	0.48	3.5–6		-.05	.30***	.08	.23***	.29***	.16*	.14*	.05	.10	.02	.77***	
13 Math grade	4.78	0.70	3.0–6		-.13*	-.05	.17**	.26***	.23***	.26***	.10	.06	.10	-.15*	.90***	.41***

Note. Sex: 0 = male, 1 = female; maternal education: 0 = no postsecondary education, 1 = postsecondary education; average grade = averaged language and math grade; α : reliability.

* $p < .05$. ** $p < .01$. *** $p < .001$.

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

1 Table 2

2 *Regression Analyses for Intelligence, Conscientiousness, and Achievement Motivation*3 *Predicting Average, Language, and Math Grades*

Step	Predictor	Average grade		Language grade		Math grade	
		β	t	β	t	β	t
1	Age	-.18	-2.76**	-.07	-1.11	-.18	-2.85**
	Sex	.10	1.57	.30	4.94***	-.05	-0.87
	Maternal education	.11	1.71	.06	1.00	.14	2.11*
	F	5.32**		9.23***		5.47**	
	R^2	.06		.10		.06	
2	Age	-.16	-2.62**	-.06	-0.93	-.17	-2.71**
	Sex	.10	1.62	.30	5.07***	-.05	-0.88
	Maternal education	.04	0.64	.00	0.06	.08	1.20
	Intelligence	.26	4.13***	.22	3.61***	.22	3.40***
	F	8.52***		10.52***		7.18***	
	R^2	.13		.15		.11	
	ΔR^2	.07		.05		.05	
3	Age	-.19	-3.17**	-.08	-1.28	-.19	-3.18**
	Sex	.02	0.39	.24	4.09***	-.12	-2.01*
	Maternal education	.06	1.02	.02	0.33	.10	1.56
	Intelligence	.25	4.17***	.22	3.59***	.21	3.39**
	Conscientiousness	.31	5.22***	.24	4.04***	.28	4.62***
	F	13.02***		12.21***		10.49***	
	R^2	.22		.20		.18	
	ΔR^2	.09		.05		.07	
4	Age	-.19	-3.30**	-.08	-1.31	-.20	-3.30**
	Sex	.13	2.05*	.32	5.11***	-.03	-0.48
	Maternal education	.04	0.73	.02	0.26	.07	1.21
	Intelligence	.21	3.64***	.19	3.23**	.17	2.81**
	Conscientiousness	.31	5.12***	.24	3.84***	.29	4.52***
	Achievement striving	.26	3.99***	.21	3.17***	.22	3.30*
	Endurance and effort	-.08	-1.15	-.09	-1.26	-.05	-0.76
	Fear of success	.07	1.11	.03	0.46	.08	1.28
	Activating test anxiety	.06	0.99	.07	1.08	.04	0.62

RUNNING HEAD: Concurrent and predictive validity for scholastic achievement

Step	Predictor	Average grade		Language grade		Math grade	
		β	t	β	t	β	t
	Inhibitory test anxiety	-.15	-2.43*	-.03	-0.49	-.19	-2.99*
	F	9.84***		7.84***		7.95***	
	R^2	.30		.25		.25	
	ΔR^2	.08		.05		.07	

4 *Note.* Sex: 0 = male, 1 = female; maternal education: 0 = no postsecondary education, 1 =

5 postsecondary education; average grade = averaged language and math grade.

6 * $p < .05$. ** $p < .01$. *** $p < .001$.

APPENDIX E: Publikationsliste

Studien

- Bünger, A., Schweizer, F. & Grob, A. Comparability of IQ test scores on group and individual level: Data from children and adolescents between 4 and 20 years. *Manuskript in Vorbereitung zur Einreichung beim Journal of School Psychology*.
- Grieder, S., Odermatt, S. D., Schweizer, F. & Grob, A. Are intelligence screenings useful? Individual and group level comparisons of screening IQ versus full-scale IQ from three different test batteries. *Manuskript in Vorbereitung zur Einreichung bei Psychological Assessment*.
- * Gygi, J. T., Hagmann-von Arx, P., Schweizer, F. & Grob, A. (2017). The predictive validity of four intelligence tests for school grades: A small sample longitudinal study. *Frontiers in Psychology*, 8, 375. doi: 10.3389/fpsyg.2017.00375
- * Schweizer, F., Grieder, S., Bünger, A. & Grob, A. Vergleich von Intelligenztestleistungen bei Monolingualen, Bilingualen und Fremdsprachigen in den Intelligence and Development Scales–2 (IDS-2). *Manuskript zur Publikation eingereicht bei Diagnostica*.
- * Schweizer, F., Gygi, J. T., Weidmann, R. & Grob, A. The concurrent and predictive validity of intelligence, conscientiousness, and achievement motivation for scholastic achievement across childhood and adolescence. *Manuskript zur Publikation eingereicht bei Personality and Individual Differences*.
- * Schweizer, F., Hagmann-von Arx, P., Ledermann, T. & Grob, A. (2018). Geschlechtsinvarianz und Geschlechtsdifferenzen in der Intelligenzeinschätzung mit den Intelligence and Development Scales. *Diagnostica*. Online-Vorveröffentlichung. doi: 10.1026/0012-1924/a000207

Testverfahren

- Grob, A. & Hagmann-von Arx, P. (2018). *Intelligence and Development Scales – 2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche*. Bern: Hogrefe. ^a

* Teil der Dissertation

^a Normierungs- und Validierungsstudien unter Mitarbeit von F. Schweizer

Stand: 13.08.2018

APPENDIX F: Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die Dissertation „Intelligenzdiagnostik im Kindes- und Jugendalter: Befunde zur Konstrukt-, differentiellen und Kriteriumsvalidität der Intelligence and Development Scales“ selbständig verfasst und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Alle wörtlichen oder sinngemäßen Zitate sind als solche gekennzeichnet. Die zur Promotion eingereichten Zeitschriftenbeiträge wurden in Zusammenarbeit mit den jeweiligen Ko-Autorinnen und Ko-Autoren angefertigt.

Basel, im August 2018

Florine Schweizer